GENOMIC ADVANCES AND CHALLENGES IN OLD AND NEW WORLD CAMELIDS

EDITED BY: Pamela Burger, Jane Collins Wheeler, Kylie Ann Munyard, Pablo Orozco-terWengel and Elena Ciani

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88966-139-8
DOI 10.3389/978-2-88966-139-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

GENOMIC ADVANCES AND CHALLENGES IN OLD AND NEW WORLD CAMELIDS

Topic Editors:
Pamela Burger, University of Veterinary Medicine, Austria
Jane Collins Wheeler, Investigación y Desarrollo de Camélidos Sudamericanos (CONOPA), Peru
Kylie Ann Munyard, Curtin University, Australia
Pablo Orozco-terWengel, Cardiff University, United Kingdom
Elena Ciani, University of Bari Aldo Moro, Italy

Citation: Burger, P., Wheeler, J. C., Munyard, K. A., Orozco-terWengel, P., Ciani, E., eds. (2020). Genomic Advances and Challenges in Old and New World Camelids. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-139-8

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 November 2018
doi: 10.3389/fgene.2018.00487

Genetic Variation in Coat Colour Genes *MC1R* and *ASIP* Provides Insights Into Domestication and Management of South American Camelids

Juan C. Marín[1], Romina Rivera[1,2], Valeria Varas[3], Jorge Cortés[1,4], Ana Agapito[1], Ana Chero[1], Alexandra Chávez[1], Warren E. Johnson[5] and Pablo Orozco-terWengel[6]*

[1] Laboratorio de Genómica y Biodiversidad, Departamento de Ciencias Básicas, Universidad del Bío-Bío, Chillán, Chile, [2] Departamento de Ciencias Básicas, Universidad Santo Tomas, Iquique, Chile, [3] Doctorado en Ciencias, Mención Ecología y Evolución, Instituto de Ciencias Ambientales & Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile, [4] Departamento de Zoología, Universidad de Concepción, Concepción, Chile, [5] Smithsonian Conservation Biology Institute, Smithsonian Institution, Washington, DC, United States, [6] School of Biosciences, Cardiff University, Cardiff, United Kingdom

OPEN ACCESS

Edited by:
Kylie Munyard,
Curtin University, Australia

Reviewed by:
Joshua Moses Miller,
Université du Québec à Montréal, Canada
Gonzalo Gajardo,
University of Los Lagos, Chile

***Correspondence:**
Juan C. Marín
jcmarin@ubiobio.cl;
dromiciops@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *14 July 2018*
Accepted: *01 October 2018*
Published: *13 November 2018*

Citation:
Marín JC, Rivera R, Varas V, Cortés J, Agapito A, Chero A, Chávez A, Johnson WE and Orozco-terWengel P (2018) Genetic Variation in Coat Colour Genes MC1R and ASIP Provides Insights Into Domestication and Management of South American Camelids. Front. Genet. 9:487. doi: 10.3389/fgene.2018.00487

The domestication of wild vicuña and guanaco by early pre-Inca cultures is an iconic example of wildlife management and domestication in the Americas. Although domestic llamas and alpacas were clearly selected for key, yet distinct, phenotypic traits, the relative patterns and direction of selection and domestication have not been confirmed using genetic approaches. However, the detailed archaeological records from the region suggest that domestication was a process carried out under significant control and planning, which would have facilitated coordinated and thus extremely effective selective pressure to achieve and maintain desired phenotypic traits. Here we link patterns of sequence variation in two well-characterised genes coding for colour variation in vertebrates and interpret the results in the context of domestication in guanacos and vicuñas. We hypothesise that colour variation in wild populations of guanacos and vicunas were strongly selected against. In contrast, variation in coat colour variation in alpaca was strongly selected for and became rapidly fixed in alpacas. In contrast, coat colour variants in llamas were of less economic value, and thus were under less selective pressure. We report for the first time the full sequence of *MC1R* and 3 exons of *ASIP* in 171 wild specimens from throughout their distribution and which represented a range of commonly observed colour patterns. We found a significant difference in the number of non-synonymous substitutions, but not synonymous substitutions among wild and domestics species. The genetic variation in *MC1R* and *ASIP* did not differentiate alpaca from llama due to the high degree of reciprocal introgression, but the combination of 11 substitutions are sufficient to distinguish domestic from wild animals. Although there is gene flow among domestic and wild species, most of the non-synonymous variation in *MC1R* and *ASIP* was not observed in wild species, presumably because these substitutions and the associated colour phenotypes are not effectively transmitted back

into wild populations. Therefore, this set of substitutions unequivocally differentiates wild from domestic animals, which will have important practical application in forensic cases involving the poaching of wild vicuñas and guanacos. These markers will also assist in identifying and studying archaeological remains pre- and post-domestication.

Keywords: alpaca, llama, vicuña, guanaco, fibre, domestication, hybridization, selection

INTRODUCTION

The first attempts at domestication coincided with the origins of agriculture some 10,000 years ago. Around that time, a global warming episode marked the end of the last ice age across the planet. More or less simultaneously in several locations of the world, a change from nomadic hunting-gathering to more-sedentary agricultural economies took place, which had a profound impact of human societies and the environment (Gepts and Papa, 2002). Once agricultural societies became more established, their domesticated plants and animals often spread from their original centres of domestication. However, continued contact with wild populations provided ample opportunities for contact and mating with the surrounding wild populations, contributing to the genetic divergence between of the domestic population from its original source population (Currat et al., 2008).

The transition from hunting and gathering to agriculture was a revolutionary inflexion point for humankind, helping support dramatic increases in human population sizes in South America, as in the rest of the world, and facilitating the emergence of modern societies (Larson and Burger, 2013; Goldberg et al., 2016). South America was the last habitable continent colonised by humans and the site of multiple domestication hotspots. Widespread sedentarism began ∼5 ka, coincident with exponential population growth promoted in part by the domestication of potatoes, common beans, peppers, groundnut, cassava, guinea pigs, llamas, and alpacas (Gepts and Papa, 2002). However, compared with the Old World, fewer vertebrate species were domesticated and these were not widely dispersed beyond their original centres of domestication. Arguably the most iconic examples occurred in the Andean high plateau, where the llama (*Lama glama*) was raised primarily as a pack animal and for its fibre and the alpaca (*Vicugna pacos*) was domesticated from the vicuña for its fine fibre.

Camelids (Artiodactyla, Camelidae) currently have both Old and New World representatives. These extant taxa originated from a common ancestor in North America 9–11 million years ago and later radiated into the Nearctic, Neotropical, and Oriental-Ethiopian regions. New World camelids originated from *Hemiauchenia*, from which descendant forms migrated into South America dispersing through either the Llanos of South America or the Cordillera de los Andes. Eventually, the Andes became the major centre for the differentiation of the currently recognised genera *Lama* Cuvier, 1800 (guanaco) and *Vicugna* Lesson, 1842 (vicuña), with guanaco inhabiting the Andes from Peru southward to Patagonia and to the Argentinean Pampa and the Chaco Region and the vicuña the higher elevations of the central Andes (Franklin, 1982; Wheeler, 1991).

Lama guanicoe, the largest South American camelid, ranges from sea level to about 3500 m between 8 and 55°S (Franklin, 1982). Two guanaco subspecies are distinguished based on subtle morphological differences. This was confirmed with well-defined genetic differentiation and subspecies designation of populations geographically separated to the northwest (*L. g. cacsilensis*) and southeast (*L. g. guanicoe*) of the central Andes plateau (Marín et al., 2008, 2013). The two vicuña subspecies are distinguished largely by body size, with the northern *V. v. mensalis* being smaller and darker in colour than *V. v. vicugna* (Marín et al., 2007a). The two other species, llama (*L. glama*, Linnaeus 1758) and alpaca (*V. pacos*, Linnaeus 1758), are domestic camelids hypothesised to be derived from *L. guanicoe* and *V. vicugna*, respectively (Kadwell et al., 2001; Wheeler, 2004; Marín et al., 2017).

Analysis of incisor morphology from faunal remains from Central Andean archaeological deposits suggests that the vicuña was first brought under human control around 7,000 years ago, leading to the domestic alpaca 1–2,000 years later (Wheeler, 1995). However, other studies have suggested that alpaca descend from guanaco, or that it is a hybrid between llama and vicuña (Hemmer, 1990). In contrast, the ancestry of llama from guanaco has been less controversial. However, it has not been established when and where domestication occurred, in part due to a paucity of archaeological evidence. Because guanaco and llama incisor morphology are identical, osteometric analyses are only able to differentiate small and large camelids among archaeofaunal records (Cartajena et al., 2007; López et al., 2012). At the molecular level, mitochondrial cytochrome b and Control Region sequences do not resolve the phylogenetic relationships among wild and domestic forms of camelids (Stanley et al., 1994; Kadwell et al., 2001; Marín et al., 2007b). In addition, various levels of hybridization among llamas and alpacas have occurred continuously (Kadwell et al., 2001), lessening the genetic distinctions between domestic forms, as well as these with their wild ancestral counterparts (**Figure 1**).

South American camelid domestication and patterns of genetic diversity were shaped initially by the needs of early Amerindians and later by shifts in breeding practises following the arrival of Europeans and the Spanish conquest (Stanley et al., 1994; Kadwell et al., 2001). It has been widely hypothesised that llamas were first selected for traits associated with producing meat for human consumption and later for their ability to carry heavy loads. In contrast, individual alpacas and their hybrid descendants were likely first selected to improve fibre quality and to produce diverse colours. Although anthropological evidence

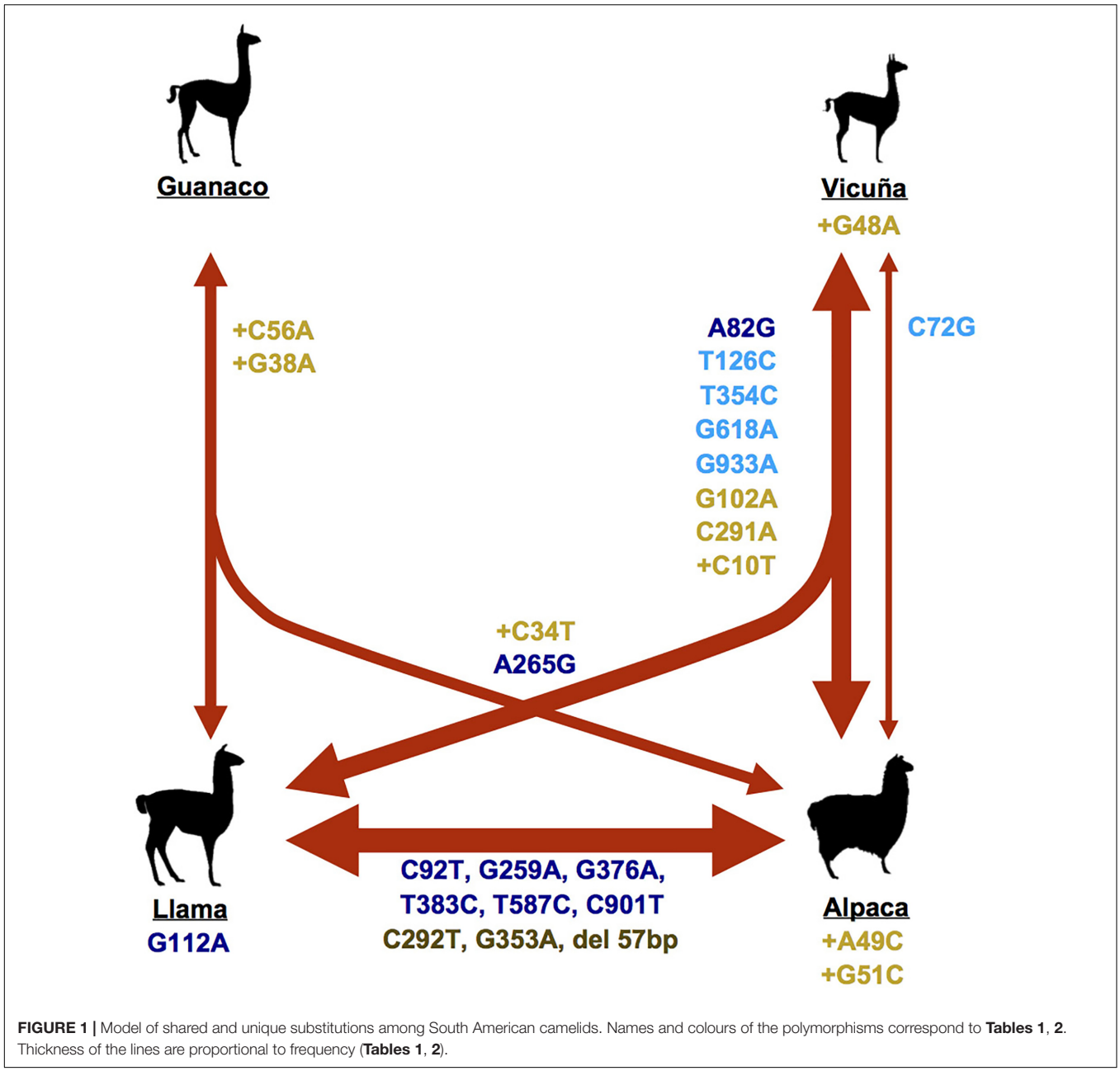


FIGURE 1 | Model of shared and unique substitutions among South American camelids. Names and colours of the polymorphisms correspond to **Tables 1**, **2**. Thickness of the lines are proportional to frequency (**Tables 1**, **2**).

has provided some insights into these processes and on the timing of domestication, including demographic and selection patterns until now the process of domestication has not been evaluated genetically.

Unlike their wild ancestors, domesticated species are often characterised by a huge allelic variability of coat colour associated genes. This variability occurs largely because artificial selection mitigates the impact of negative pleiotropic effects that otherwise are linked with certain coat colour variants in wild populations (Dong et al., 2015). But artificial selection during domestication is different from a standard selective sweep of a new strongly beneficial variable because artificial selection acts on standing genetic variation, which may have been neutral of even selected against before domestication. Therefore, the fixation of a beneficial allele does not always wipe out DNA variation in the surrounding region and the amount by which variation is reduced largely depends on the initial frequency of the beneficial allele. Recent studies demonstrate that selection for coat colour phenotypes started at the beginning of domestication. Although more than 300 genetic loci and more than 150 coat colour-associated genes have been described, the genetic pathways that determine coat colouration are still poorly understood. However, a few conserved genes have consistently been implicated in coat colour patterns and genes with unique contributions to colour patterns continue to be identified (Cieslak et al., 2011).

Here, we characterised genetic patterns of modern South American camelids in two well-studied coat colour genes, the Melanocortin 1 receptor (*MC1R*) and Agouti-signalling peptide (*ASIP*). Colour patterns of mammalian skin and hair are determined by a relatively small number of genes. These genes have been classified into those active in the development, differentiation, proliferation, and migration of melanocytes and those acting directly on the pigment synthesis. *MC1R* and *ASIP* are two of the best-described genes involved in the synthesis of pigment. *MC1R* encodes the melanocortin receptor that, when coupled to G-proteins, stimulates the production of eumelanin and is responsible for dark colours in melanocytes. *ASIP* produces the agouti signalling protein that acts as an antagonist of *MC1R* by annulling the (∝-MSH) action, thus favouring the production of pheomelanin which produces light colours in melanocytes.

MC1R of most vertebrates shares a characteristic allelic hierarchy, where dominant allele (E) produces dark pigment and the recessive allele (e) produces light pigments. Similar mutational patterns have been described in an array of species, including domestic dog (Schmutz et al., 2003), pig (Kijas et al., 1998), and horse (Marklund et al., 1996). In contrast, the dominant *ASIP* allele (A) produces a yellow–red colour pattern while the recessive allele (a) is linked with uniform black. Loss-of-function substitutions in *ASIP* linked with eumelanic phenotypes have been described in several species (Rieder et al., 2001; Eizirik et al., 2003; Kerns et al., 2004; Royo et al., 2008). Although *MC1R* is primarily responsible for determining which pigment type is produced, both *MC1R* and *ASIP* can act locally, causing non-uniform colouration in different regions of the body (Cieslak et al., 2011). Dominant black alleles and putative recessive alleles from the *MC1R* gene have been identified in sheep (Våge et al., 1999; Fontanesi et al., 2011) and goats (Fontanesi et al., 2009). Copy number variation of *ASIP* has also been associated with light and dark coats in goats and sheep (Norris and Whan, 2008; Fontanesi et al., 2011; Dong et al., 2015). In alpaca and llama, several polymorphic sites have been identified on *MC1R* and *ASIP* (Powell et al., 2008; Bathrachalam et al., 2011; Feeley et al., 2011; Daveiro et al., 2016). However, these substitutions have yet to be evaluated in their wild counterparts.

Although guanacos and vicuñas have very homogeneous coat colour patterns, their domestic counterparts exhibit a remarkable range of colours and patterns. However, understanding the process by which these coat colour patterns were selected has been complicated by a lack of molecular markers that discriminate among the four species and by evidence of significant historic gene flow between domestic forms during different time periods. Here we hypothesise that colour variation in wild populations of guanacos and vicunas was strongly selected against. In contrast, variety in coat colour variation in alpaca was strongly selected for, while coat colour variants in llamas were of less economic value thus were under less selective pressure. More specifically, our goal was to identify and quantify patterns of molecular variation in *MC1R* and *ASIP* in the four species of CSA and to determine if there are specific polymorphisms or genetic patterns that are associated with certain colours or that are able to discriminant between the wild and domestic forms of these iconic and economically important South American species. The implications of the results will be discussed in the context of more-recent selection for fibre colour and quality by breeders worldwide.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

Between 1994 and 2016 samples were collected from 82 guanacos and 89 vicuñas from Peru, Bolivia, Argentina, and Chile, encompassing the entire range of distribution of each species (**Supplementary Table 1**). These samples were complemented with a collection of 89 llamas and 84 alpacas from Ecuador, Peru, Argentina, and Chile were collected and analysed (total dataset 344 samples; **Supplementary Table 1**). Samples comprised skin (n = 4), muscle (n = 3), and blood (n = 337), and were stored at −70°C in the Laboratorio de Genómica y Biodiversidad, Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad del Bío-Bío, Chillán, Chile or at CONOPA in Lima, Peru. Total genomic DNA was extracted from blood using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI, United States) following the manufacturer's instructions. DNA from skin and muscle samples was purified using proteinase K digestion and a standard phenol-chloroform protocol (Sambrook et al., 1989). Llamas and alpacas were grouped into six colour groups (black, dark brown, light brown, grey, white, and wild). Sampling permits are provided in the acknowledgements section.

Amplification and Sequencing of *ASIP* and *MC1R*

Polymerase chain reaction (PCR) primers were designed to amplify the complete *MC1R* coding, following Feeley and Munyard (2009) and the *ASIP* coding region and intronic portions were amplified using primers designed described in Feeley et al. (2011) (**Supplementary Table 2**). Amplification was performed in a 50 l L reaction volume with ∼30 ng genomic DNA, 1× PCR buffer (8 mM Tris–HCl (pH 8.4); 20 mM KCl (Invitrogen Gibco, Life Technologies), 2 mM MgCl2, 25 l M each of dNTP, 0.5 lM each primer and 0.1U/l Taq polymerase (Invitrogen Gibco, Life Technologies®). All PCR amplifications were performed in a Veriti® thermal cycler (Applied Biosystems, Paisley, United Kingdom) with cycling conditions as follows: initial denaturation at 95°C for 10 min, followed by 30–35 cycles of 94°C for 30 s, annealing for 30 s (**Supplementary Table 2**) and 72°C for 60 s, and a final extension of 72°C for 5 min. PCR products were purified using the GeneClean Turbo for PCR Kit (Bio101) following the manufacturer's instructions. Products were sequenced in forward and reverse directions using BigDye chemistry on an ABI Prism 3100 semi-automated DNA analyser. The reactions were carried out in a 10 μl volume containing approximately 100 ng of purified DNA, 1 μl of either forward or reverse primer and 2 μl of BigDye Terminator Kit version 3.1 (PerkinElmer). Sequence reactions were visualised using an ABI-3100 sequencer (Perkin Elmer Applied Biosystems).

Sequence Analysis

Single nucleotide polymorphisms (SNPs) were identified by sequence alignment using Geneious[1] and were confirmed by resequencing the whole fragment in the opposite direction. *MC1R* and *ASIP* gene sequences were deposited in GenBank with accession numbers MH596009–MH596352 and MH596353–MH596692). Aligned sequence data for each gene separately were imported into DNASP 5.0 software (Librado and Rozas, 2009) to analyse haplotype diversity and nucleotide diversity. The gametic phase of each haplotype was determined with the software BEAGLE Version 3.3.1 (Browning and Browning, 2007). The genealogical relationship of *MC1R* and *ASIP* haplotypes was described with a haplotype network using the uncorrected median-joining values in Splits Tree4 V 4.14.6 (Huson and Bryant, 2006).

Detection of Loci Under Selection and Association With Phenotypes

To assess if SNP loci were under selection in the four species of South American camelids, we used BayeScan (Foll and Gaggiotti, 2008), which decomposes locus-population F_{ST} coefficients into a population-specific component (beta) and a locus-specific component (alpha) with positive alpha values indicating diversifying selection. BayeScan was run for each gene with default values under the codominant marker model and assuming two populations, i.e., one with all wild animals and the other with the domestic ones. We also used hapFLK (Fariello et al., 2013) to detect haplotypes under selection, grouping the haplotypes observed in each gene into two populations as done for BayeScan and testing if the haplotype frequencies fit a neutral model. We tested whether the distribution of dN and of synonymous substitutions (dS) in both genes (**Tables 1**, **2**) was the same in the wild and domestic species using chi-square tests calculated for each substitution type separately, and tested if the number of alleles at each polymorphism was the same in the two wild species vs. the two domestic species using Cochran–Mantel–Haenszel tests. Associations between genotypes and coat colour phenotypes were determined using haplo.stats (Sinnwell et al., 2007). Haplo.stats uses an Expectation-Maximisation algorithm to first determine the frequency of haplotypes for each individual and then provides a haplotype-specific score to test for significant differences among cases and controls using a chi-square distribution with degrees of freedom equal to the number of inferred haplotypes in each haplotype block. For these analyses we assessed the correlation of several colour phenotypes in each of the wild and domestic species, however, considering each comparison as a binary system (e.g., animals with black phenotype vs. all animals with a different phenotype). The phenotypes tested were wild (N = 171), black (N = 15), dark brown (N = 15), light brown (N = 43), grey (N = 4), and white (N = 35). Only categories with haplotype frequencies greater than 1% were assessed.

[1] www.geneious.com

RESULTS

MC1R

A fragment of 954bp of the *MC1R* gene was amplified in all samples. Fourteen single nucleotide polymorphisms (SNPs) were identified in this fragment, nine of which presented non-synonymous substitutions (**Supplementary Figure 1** and **Table 1**). Overall, we described 74 haplotypes (**Supplementary Table 3**) and total haplotype (h) and nucleotide (p) diversity of 0.875 and 0.0037, respectively (**Supplementary Table 4**). Of the 14 SNPs detected, one (354 T > C) is common in pig (Adeola et al., 2017) and another (SNP 901C > T) has been observed in Old World camelids (Almathen et al., 2018). Substitutions in *MC1R* related with coat colour have also been documented in several domestic mammals (Schmutz et al., 2003; Wang et al., 2013; Adeola et al., 2017; Wu et al., 2017). Our analyses also identified two polymorphisms (c.72 C > G and c.265 A > G) that have not been previously reported in alpacas or llamas but which are present in vicuñas and alpacas. Additionally, one new substitution (c.265 A > G) was detected in low frequencies in the four species (**Table 1**).

To elucidate the genealogical relationship between the South American Camelids (SAC) *MC1R* haplotypes, we drew a network plot for 39 haplotypes (**Figure 2**). Overall, a division between the haplotypes in wild *Vicugna* and wild *Lama* species was observed, however, the domestic species shared haplotypes with each other and with their wild relatives. A greater separation between vicuñas from the south and north is evidenced, with the northern ones being closer and sharing more haplotypes with the domestic species, especially alpacas. Moreover, the network did not exhibit a clear genetic partition between llama and alpaca with six haplotypes shared between them, and with more haplotypes shared between the domestic species and *Lama* haplotypes than with *Vicugna* haplotypes.

Translation of the *MC1R* sequence revealed an open reading frame of 318 amino acids. Five of the fourteen SNPs were synonymous substitutions (L24L, D42D, N118N, L206L, and E311E) while the remaining nine resulted in amino acid substitutions (T28A, T31M, V38M, V87M, M89V, G126S, M128T, F196S, and R301C). A comparison between the number of non-synonymous (dN) and synonymous (dS) substitutions among the four species and among the wild and domestics showed an excess of dN substitutions in the domestic species (p = 0.0258, and p = 0.002, respectively) while no significant differences in the number of dS substitutions (p = 0.438, and p = 0.4126, respectively). A significant difference in the number of dN and dS among species (p = 0.01655) and between the wild and domestics (p = 0.0079) was found (**Table 3**). Of the fourteen substitutions detected in *MC1R*, seven were absent in wild SACs and were found only in domestic species; all of these were non-synonymous changes (**Figure 1** and **Table 4**).

A significant difference in the number of alleles in the wild vs. domestic species comparison was observed with the Cochran–Mantel–Haenszel test ($p < 0.001$; **Table 4**) for thirteen of the 14 *MC1R* substitutions. All of the non-synonymous substitutions with non-significant p-values occurred at very low frequencies

TABLE 1 | Single nucleotide polymorphism (SNP) variation detected in *MC1R* of wild and domestics South American camelids.

Polymorphism	Amino acid change	Location	Amino acid effect	Type of substitution	Guanaco	Vicuña	Llama	Alpaca
C72G	Leu	c. 72	N/A	**dS**	0 (0)	34 (7)	0 (0)	0 (1)
A82G	Thr/Ala	c. 82	Polar to non-polar	**dN**	0 (0)	0 (1)	3 (5)	10 (26)
C92T	Thr/Met	c. 92	Polar to non-polar	**dN**	0 (0)	0 (0)	3 (12)	0 (3)
G112A	Val/Met	c. 112	Polar to polar	**dN**	0 (0)	0 (0)	0 (13)	0 (0)
T126C	Asp	c. 126	N/A	**dS**	0 (0)	2 (2)	2 (6)	9 (29)
G259A	Val/Met	c. 259	Non-polar to polar	**dN**	0 (0)	0 (0)	24 (24)	30 (30)
A265G	Met/Val	c. 265	Non-polar to polar	**dN**	0 (1)	8 (8)	0 (1)	0 (2)
T354C	Asn	c. 354	N/A	**dS**	0 (0)	4 (4)	1 (1)	9 (24)
G376A	Gly/Ser	c. 376	Polar to polar	**dN**	0 (0)	0 (0)	19 (24)	27 (30)
T383C	Met/Thr	c. 383	Non-polar to polar	**dN**	0 (0)	0 (0)	3 (24)	0 (7)
T587C	Phe/Ser	c. 587	Non-polar to polar	**dN**	0 (0)	0 (0)	1 (1)	0 (2)
G618A	Leu	c. 618	N/A	**dS**	0 (0)	8 (8)	4 (3)	10 (16)
C901T	Arg/Cys	c. 901	Changed to polar	**dN**	0 (0)	3 (0)	5 (0)	14 (8)
G933A	Glu	c. 933	N/A	**dS**	13 (0)	70 (3)	7 (2)	24 (1)

Number of homozygote (heterozygote) individuals for each polymorphic site in guanacos (N = 82), vicuñas (N = 89), llamas (N = 89), and alpacas (N = 84). Type of substitution (dS indicates synonymous substitutions and dN indicates non-synonymous substitutions).

TABLE 2 | Single nucleotide polymorphism variation in *ASIP* of wild and domestic South American camelids.

Polymorphism	Amino acid change	Location	Amino acid effect	Type of substitution	Guanaco	Vicuña	Llama	Alpaca
G102A	Gly	c. 102, Exon 2	N/A	**dS**	0 (0)	3 (19)	0 (7)	11 (32)
+C34T	N/A	+34, Exon 3	N/A	**dS**	0 (1)	6 (23)	0 (4)	9 (27)
+G48A	N/A	+48, Exon 3	N/A	**dS**	0 (0)	0 (15)	0 (0)	0 (0)
+A49C	N/A	+49, Exon 3	N/A	**dS**	0 (0)	0 (0)	0 (0)	2 (0)
+G51C	N/A	+51, Exon 3	N/A	**dS**	0 (0)	0 (0)	0 (0)	2 (0)
+C56A	N/A	+56, Exon 3	N/A	**dS**	28 (21)	0 (1)	38 (16)	29 (19)
C291A	Thr	c. 291, Exon 4	N/A	**dS**	0 (0)	12 (26)	1 (3)	11 (25)
C292T	Arg/Cys	c. 292, Exon 4	Basic to polar	**dN**	0 (0)	0 (0)	3 (21)	5 (23)
del 57 bp	Cys109-Arg127del	325-381, Exon 4	Displacement	**N/A**	0 (0)	0 (0)	23 (17)	6 (24)
G353A	Arg/His	c. 353, Exon 4	Basic to polar	**dN**	0 (0)	0 (0)	1 (13)	16 (18)
+C10T	N/A	+10, Exon 4	N/A	**dS**	0 (0)	9 (28)	0 (4)	13 (21)
+G38A	N/A	+38, Exon 4	N/A	**dS**	30 (17)	0 (1)	40 (9)	21 (25)

Number of homozygote (heterozygote) individuals for each polymorphism in guanacos (N = 82), vicuñas (N = 89), llamas (N = 89), and alpacas (N = 84). Type of substitution (dS indicates synonymous substitutions and dN indicates non-synonymous substitutions).

TABLE 3 | Number of non-synonymous (dN) and synonymous (dS) substitutions in *MC1R* and *ASIP* among species and among wild and domestic South American camelids.

Species	Number of *MC1R* substitutions		Number of *ASIP* substitutions		Total number of substitutions	
	dN	dS	dN	dS	dN	dS
Wild	4	6	0	10	4	16
Domestic	18	9	6	14	24	23
χ^2 *p*-values	**0.002838**	0.4386	**0.01431**	0.4142	**0.0001571**	0.2623
Guanaco	1	1	0	3	1	4
Vicuña	3	5	0	7	3	12
Llama	9	4	3	6	12	10
Alpaca	9	5	3	8	12	13
χ^2 *p*-values dN vs. dS species by species	**0.01656**		**0.002167**		**0.02358**	
χ^2 *p*-values wild vs. domestic for each genus	**0.00796**		**0.000685**		**0.01087**	
χ^2 *p*-values wild vs. domestic within substitution type	**0.02588**	0.4126	0.1116	0.5062	**0.002222**	0.1718

*Numbers in bold denote statistically significant values (***$P < 0.05$) using χ^2 test.*

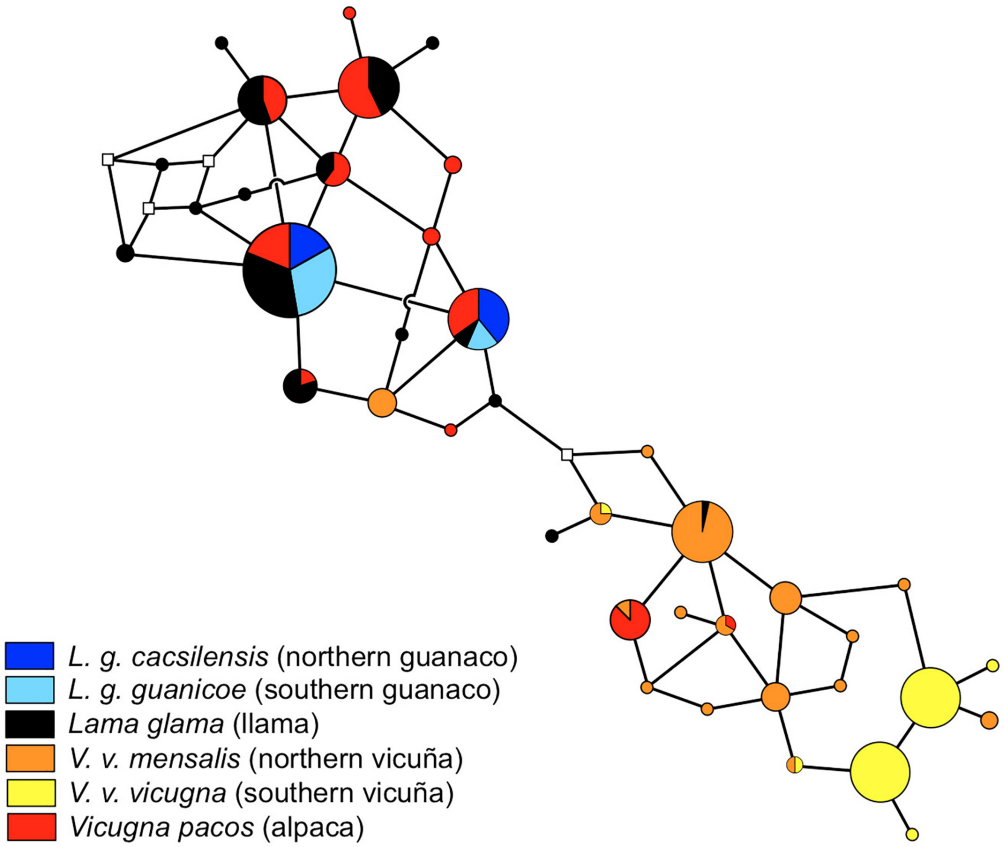


FIGURE 2 | Median-joining network *MC1R* gene genealogy of South American camelids using split decomposition using SplitsTree, version 4.0 (Huson and Bryant, 2006). Circle sizes correspond to haplotype frequencies. The sphere colours correspond to different taxa.

(**Table 4**). Only 8 of the llamas (8.9%) and 3 of the alpacas (3.6%) had none of the seven substitutions that were unique to the domestic animals. Two of these exclusive domestic substitutions (G259A and G376A) were observed at high frequency (more than 50% of animals) in llamas and alpacas. G259A was present in 52.8% of the llamas and 73.2% of the alpacas; G376A was present in 48.3% of the llamas and 69.5% of the alpacas (**Table 4**).

Two substitutions (C72G and A265G, synonymous and non-synonymous, respectively) were observed in the southern vicuñas and were very rare in the domestic species (four heterozygous individuals) supporting the hypothesis that alpacas are derived from the northern vicuñas (*V. v. mensalis*). This is further supported by the synonymous substitution G618A which is almost exclusive in northern vicuñas, and is also observed in some alpacas and llamas, but which occurs at very low frequency in the southern vicuñas (**Supplementary Table 5**).

ASIP

From the same individuals, we obtained the complete coding sequence of *ASIP* and assessed patterns of variation in a 402-bp gene fragment consisting of 159, 66, and 177 bp of Exons 2, 3, and 4, respectively. In addition, 229-bp of intronic sequence downstream of each exon were also analysed (**Supplementary Figure 2** and **Table 2**). Combined, these constituted 49 haplotypes (**Supplementary Table 6**) wit total haplotype (h), and nucleotide (p) diversities of 0.782 and 0.004, respectively, for exons and introns combined and nucleotide (p) diversities of 0.0052 and 0.0145 for the exons and introns and haplotypic (h) diversities of 0.433 and 0.709, respectively (**Supplementary Table 7**).

The *ASIP* network depicted a relatively simple genealogy with a predominant haplotype shared among all taxa. Unlike the genealogy of *MC1R*, *ASIP* does not show a separation between the *Vicugna* and *Lama* genera, however, ten low frequency and related to each other haplotypes are exclusive to *Vicugna* (**Figure 3**).

ASIP sequences had an open reading frame of 134 amino acids. Two of the 12 substitution cause amino acid changes (C292T and G353A) and one deletion (at position p.C109-Rdel19) is predicted to result in 19 of the 25 amino acids being absent from the mature protein (**Table 2**). This deletion would likely eliminate the two beta sheets and the R-F-F-motif from the agouti functional domain, which are considered to be essential to interact with ∝-MSH. The other seven observed substitutions are outside the coding region (+C34T, +G48A, +A49C, +G51C, +C56A, C291A, +C10T, and +G38A) or are synonymous changes (G102A).

TABLE 4 | Polymorphisms with strongest association signals for domestics species.

Species	Non-synonymous substitutions of *MC1R*																Synonymous substitutions of *MC1R*									
	A82G		C92T		G112A		G259A		A265G		G376A		T383C		T587C		C901T		C72G		T126C		T354C		G618A		G933A	
	A	G	C	T	G	A	G	A	A	G	G	A	T	C	T	C	C	T	C	G	T	C	T	C	G	A	G	A
Guanaco	164	0	82	0	164	0	164	0	163	1	164	0	164	0	164	0	164	0	164	0	0	0	0	0	164	0	0	0
Llama	178	11	160	18	165	13	106	72	177	1	116	62	148	30	175	3	168	10	178	0	168	10	175	3	167	11	164	14
Vicuña	177	75	89	0	178	0	178	0	100	78	178	0	178	0	178	0	172	6	103	75	2	176	4	174	94	84	53	125
Alpaca	122	46	165	3	168	0	75	93	166	2	81	87	161	7	166	2	130	38	167	1	121	47	124	44	129	39	115	53
p-value	9.59E-14		0.002738		0.001192		2.58E-47		7.29E-20		3.87E-42		2.32E-09		0.078		5.72E-09		4.13E-20		2.37E-42		4.32E-44		0.000479		5.66E-14	

Cochran–Mantel–Haenszel association analysis of non-synonymous and synonymous substitutions of MC1R. Association statistics are reported for 14 of the 15 substitutions.

Nine substitutions were synonymous, one in the Exon 2 (G34G) and one in the Exon 4 (T97T). The others were detected downstream of Exon 3 and 4. Only 2 non-synonymous (dN) substitutions and one deletion of 57 bp were detected in Exon 4. This deletion includes nucleotides 325–381 and would likely mask the polymorphic site G353A (c.325–381 del 57). Additionally, nine synonymous substitutions were detected, seven of which were upstream from exons 3 and 4 and two were in exons 2 and 4 (**Supplementary Figure 2**). Of the 12 SNPs detected, three have been observed in pig too (Wu et al., 2017). Seven of these substitutions are reported here for the first time in alpacas, llamas, and vicuñas. Only three of these were present in all species and all were synonymous substitutions (**Tables 2**, **5**).

The difference in the number of *ASIP* dN substitutions among wild and domestics was significant (p = 0.01431), while that between dS substitutions was not (p = 0.4142). A significant difference in the number of dN and dS among species (p = 0.00217) and between the wild and domestics (p = 0.0007) was found (**Table 3**). In contrast, both wild and domestic animals present similar levels of dN (p = 0.11) and dS (p = 0.51) substitutions (**Table 3**). Of the twelve substitutions detected in *ASIP*, five were absent in the wild, with two of these corresponding to non-synonymous changes (**Figure 1** and **Table 4**).

Significant difference in the allelic count was observed for 7 out 12 substitutions (Cochran–Mantel–Haenszel $p < 0.001$) (**Table 5**). Five of the substitutions detected in *ASIP* (C292T, G353A, del 57, A+49C, and G+51C) were found only in domestic animals (llamas and/or alpacas). Two of these (C292T and G353A) change the amino acid and the deletion produces a shift in the reading frame. These exclusive domestic substitutions were very common (over 44% of animals), especially in alpacas. Importantly, two dS substitutions of the *ASIP* gene (+C56A and +G38A) discriminated northern (*L. g. cacsilencis*) from southern guanacos (*L. g. guanicoe*). These substitutions were frequently observed in homozygosis in llamas, being observed in 72.7 and 57.1% of llamas and 50.0 and 65.7% of alpacas, respectively, supporting the hypothesis that llamas are derived from the northern guanacos (**Supplementary Table 8**).

Detection of Loci Under Selection and Association Analysis

Comparison among wild and domestic species using BayeScan and hapFLK did not identify SNPs under selection. Because of the shared genetic variation due to hybridisation we focused these analyses only on those animals with ≥70 and ≥90% assignment coefficient to their respective group. The BayeScan method didn't detect outlier loci (FDR = 0.05), however, seven *MC1R* SNPs showed positive alpha values, while only one SNP in *ASIP* showed a positive alpha value. F_{ST}-values for these SNPs ranged between 0.44 and 0.45 for *M1CR* and between 0.26 and 0.28 for *ASIP*. The hapFLK analysis resulted in p-values larger than 0.05 in all tests (**Supplementary Figures 3A,B**, respectively).

There was a strong association linking *MC1R* haplotypes 1, 2, and 3 and the wild species status (p = 3.3e-7, p = 5.3e-10, and p = 6.7e-12, respectively). Specifically, haplotype 2 was

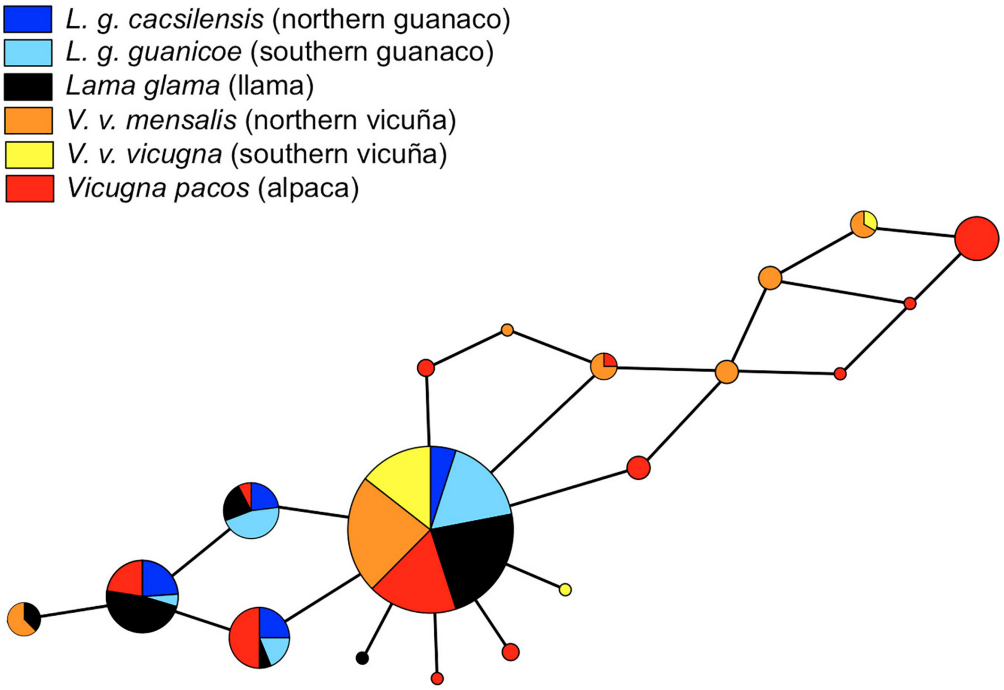


FIGURE 3 | Median-Joining network of *ASIP* gene genealogy of South American camelids using split decomposition using SplitsTree, version 4.0 (Huson and Bryant, 2006). Circle sizes correspond to haplotype frequencies. The sphere colours correspond to different taxa.

associated with northern and southern guanacos, haplotypes 1 and 3 with northern vicuñas and haplotype 3 was associated with southern vicuñas (p = 3.9e-32). Additionally, there was an association of domestic species with haplotypes 4, 5, 6, and 7 (p = 7.9e-4, p = 1.7e-4, p = 1.8e-5, and p = 1e-21, respectively). Haplotypes 4 and 5 were linked with alpacas (p = 1.2e-10 and p = 2.8e-12, respectively) and haplotype 6 with llamas (p = 1.7e-10). Haplotype 7 was highly associated with both domestic species (**Supplementary Table 9**).

With *ASIP* there was an association between haplotype 2 and the white colour in alpacas (p = 1.6e-2). Also, haplotype 3 was associated with the variety of light colours (light brown, cream, and white; p = 2.3e-6, p = 7.3e-4, and p = 3.1e-3, respectively). There was no clear association of haplotype 4 with any colour, perhaps because this gene has been mainly linked in other species with dark brown, light brown, cream, and white colours. Other associations were not as robust. However, haplotype 1 was strongly linked with wild species (p = 1.9e-40), without discriminating between subspecies. Haplotypes 2, 3, and 4 had a high association with alpaca and llama (p = 5.5e-8, p = 8.1e-12, and p = 9.8e-16, respectively) (**Supplementary Table 10**).

DISCUSSION

Archaeological records suggest that the domestication of llamas and alpacas involved very active and directed management practises. This process likely applied strong selection pressure to increase the frequency of desired phenotypic traits, and would have fixed desired (beneficial) phenotypes in the founder populations of domesticated species relatively rapidly. These fixation events differ from the fixation of an advantageous mutant in a natural population, in that artificial selection in a domestication event is likely act on standing genetic variation of the ancestral population which was neutral or nearly neutral before domestication (e.g., Kijas et al., 2012). Contrastingly, through the domestication process it is also possible that selection becomes relaxed as genetic variants under purifying selection in the wild may not be selected against in the protected domestic environment (e.g., Alberto et al., 2018).

South American camelids offer a unique opportunity to study the evolution and domestication of a group of large mammals, particularly since the domestication of llamas and alpacas is relatively well documented and resulted from the selective breeding of two wild species, the guanaco and the vicuña, that still persist. The data generated here supports to a large extent these relationships as shown by our SplitsTree networks for each gene, which reveal similar, yet somewhat contrasting results. Both genes depict the links of llamas with both guanaco subspecies and of the northern vicuna subspecies with the alpaca, however, the resolution differentiating vicuña from guanacos is higher for *MC1R* than *ASIP*. With *ASIP*, the most common haplotype occurs in each of the groups and the majority of the minor haplotypes occurred in alpaca and the northern vicuña. These results are in accordance with the archaeological and cultural records that indicate that llamas were initially bred from guanacos as a source of meat and later for docility and carrying heavy loads and that alpaca were bred from vicuñas for docility and their

TABLE 5 | Polymorphisms with strongest association signals for domestic species.

Species	No-synonymous substitutions of *ASIP*										Synonymous substitutions of *ASIP*													
	C292T		G353A		del 57 bp		G102A		+C34T		+G48A		+A49C		+G51C		+C56A		C291A		+C10T		+G38A	
	C	T	G	A	+	–	G	A	C	T	G	A	A	C	G	C	C	A	C	A	C	T	G	A
Guanaco	162	0	162	0	162	0	160	0	161	1	160	0	162	0	162	0	85	77	162	0	162	0	85	77
Llama	139	27	161	5	113	63	169	7	172	4	176	0	176	0	176	0	74	102	161	5	162	4	77	89
Vicuña	176	0	176	0	178	0	151	25	141	35	161	15	178	0	176	0	175	1	122	50	126	46	171	1
Alpaca	107	33	90	50	132	36	112	54	123	45	168	0	164	4	164	4	91	77	93	47	93	47	73	67
p-value	2.12E-17		2.44E-18		5.42E-26		7.03E-06		0.094		0.000318		0.118		0.12		4.30E-15		0.197		0.104		1.75E-12	

Cochran–Mantel–Haenszel association analysis of non-synonymous and synonymous substitutions of ASIP. Association statistics are reported for 7 of the 12 substitutions.

fine fibre of different colours, becoming essential animals for lifestyle and economy of the Andean people. However, our results demonstrate conclusively that following domestication, llamas and alpacas did not remain genetically isolated and that hybridization between both alpacas and llamas would have occurred regularly, and rare haplotypes, perhaps associated with colour variation, were more-often retained in alpacas.

Overall, we find that the domestic species seem to harbour a substantially higher level of genetic variation than their wild counterparts, and the domestics' variation has been modified by the effect of artificial selection. In particular, natural selection in the wild is expected to have kept variation in the genes involved in coat colour to a minimum favouring a phenotype of higher fitness, as has been shown for other domestic species (e.g., Gratten et al., 2007; Chessa et al., 2009; Ludwig et al., 2009; Cieslak et al., 2011; Li et al., 2014). However, under domestication, it is expected that variants otherwise removed by selection may have been selected for by humans as phenotypes of interest (Alberto et al., 2018). This is supported by the similarity in the frequency of dN and dS substitutions in a relatively high proportion of SNPs found unique to llamas and alpacas (12 of 26), as well as the genetic variation in wild animals that is also present in domestic animals (**Figure 1**). It is possible that the substitutions for the coat colour produced in alpacas or vicuñas, particularly those that produce amino acid changes and therefore possibly different phenotypes, were incorporated rapidly into the llamas, possibly with the use of male llamas that were crossed with females alpacas, as previously described with paternal and maternal markers (Marín et al., 2017).

The limited number of polymorphic sites among guanacos across their broad distribution could indicate that there is stronger selection pressure for more-homogeneous colour phenotypes in guanacos, perhaps related with the broad and heterogeneous habitats that guanacos inhabit ranging from the coast (sea-level) to 3,500 m above sea level. In contrast, the large number of shared substitutions among vicuñas and both domestic species (**Figure 1**) could be an indication that the vicuña underwent strong selection for their fibre quality at a later stage of domestication, eventually being the South American camelids with the smallest diameter of fibre and one of the thinnest fibres among mammals, with fibre diameter of 12.52 ± 1.52 μm (Carpio and Solari, 1982). Similarly, sheep in Eurasia went through at least two successive periods of selection, initially for its meat and 4,000–5,000 years ago for its milk and fine wool (Kijas et al., 2012). Contrary to what is currently common thought, it is possible, that the vicuñas were domesticated after the guanacos when the communities had already satisfied the need for food and became more focus on better clothing. Except for the 57 bp deletion of Exon 4 of the *ASIP* gene, which appears more frequently in llamas, the two substitutions with the strongest signals of selection in *MC1R* gene (G259A and G376A), as well as the other detected substitutions, are more common in alpacas. This pattern that may be a evidence of strong artificial selection (e.g., line breeding) or relaxation of the selection against colour types as fibres of different colours in the alpacas could have generated a great variety of colours that later would have

been bred into the llamas as an inevitable consequence of interbreeding.

During the height of the Inca empire, management of domestic and wildlife populations of camelids were likely under relative strict control of central authorities. Agricultural practises were historically maintained by oral tradition and within a specialised group called the *yana*. Animals were segregated and bred based on their desired characteristics and breeding records were recorded with an information storage instrument called the *quipu* (Wheeler et al., 1992). It is not clear to what extent the pre-Columbian alpacas were selected for their fibre diameter or for the purity of colour. However, most of this sophisticated management system and accumulated wisdom would have been lost after the Spanish conquest and as camelids were replaced by traditional sheep breeders, whose main objective was to increase yield (weight) of fibre instead of specific quality or colour. Archaeological remains supporting this scenario have been found in El Yaral, Peru (Wheeler et al., 1995).

The genetic markers and patterns of variation described here have the potential to help identify and protect possible relict populations of pre-conquest alpaca and llama phenotypes (Wheeler et al., 1995). The probable role of hybridization in the evolution of today's llamas and alpacas is unknown, but it is possible that the 16th century introduction of sheep, cattle, and horses into the region also led to a breakdown in controlled breeding accompanied by extensive hybridization produced by events of the conquest (Wheeler et al., 1995). A study similar to this one, with candidate genes for fibre diameter, will be necessary to test this hypothesis and to be able to develop a full array genetic markers and management tools to assist in improving herds and restoring the local textile industry.

The presence of different coat colours is a typical characteristic of domestic species, it often constitutes one of the phenotypes selected early during domestication in mammals (Fang et al., 2009; Ludwig et al., 2009; Cieslak et al., 2011; Dong et al., 2015), and reflects phenotypes that under wild conditions may be of lower fitness. Consistent with this hypothesis, we find that the number of dN substitutions is significantly lower in wild SACs in comparison with domestic SACs.

Although we know that there is gene flow among the four species, especially between domestic animals or between wild and their domestic derivatives, almost all non-synonymous substitutions were not observed in wild individuals suggesting strong selection against them, or their phenotypes, in guanacos and free-living vicuñas. It is possible then that the selection of special colours in alpacas occurred mainly following domestication, rather than through the capture of albino or melanic vicuñas as has characterised impala, bontebok and wildebeest captive breeding in Southern Africa (Russo et al., 2018).

Notably, the A82G substitution dN of the *MC1R* is present in all the vicuñas and none of the guanacos, clearly differentiating the *Lama* from the *Vicugna*. Two dS substitutions of the same gene (T126C and T354C) show this same distinctive pattern between the two genera. Only some llamas and a greater number of alpacas present the substitution that changes Threonine by Leucine. It is possible that this polymorphism is related to the difference in colour between guanacos and vicuñas, in which the vicuña, with a dark cinnamon colour, could be expressing more pheomelanin than guanacos, whose eumelanin expression could be responsible for the more dark, reddish brown patterns in the southern populations (*L. g. guanicoe*) compared with the lighter brown with ochre yellow tones in the northern variety (*L. g. cacsilensis*) (Wheeler, 2012). The A265G substitution, on the other hand, although it is also scarcely present in guanacos, llamas and alpacas (with 1, 1 and 2 heterozygous individuals, respectively), is a diagnostic substitution for southern vicuñas. This substitution may also be responsible for the tonality differences observed in the southernmost vicuñas, which typically are more yellowish, and which could be related with subtle differences in expression of pheomelanic pigments. Similarly, the dS substitutions C72G and G618A, detected in the *MC1R* gene, together with A265G, are also diagnostic polymorphisms that unequivocally differentiate the two-vicuña subspecies. Interestingly, these three substitutions are heterozygous in individuals inhabiting the contact zone between *V. v. mensalis* and *V. v. vicugna*. Similarly, a less strong signal from the dS + C56A and + G38A substitutions distinguish the more northern from the southernmost guanacos. Since the association between guanaco and llama is not as strong with these substitutions, this may be evidence that guanaco domestication was not primarily motivated by selection for fibre quality or colour.

Finally, although the detected polymorphisms here do not distinguish the domestic species from each other, for the first time, polymorphic sites have been described that distinguish wild South American camelid species from their domestic derivatives. These guanacos and vicuñas haplotypes diagnosed, combined with 10 dN substitutions in llamas and alpacas, will make it possible to distinguish wild from domestic camelid samples. This result will have important implication for forensic applications, including the control of wildlife trade and the conservation of wild species. Application of these techniques and markers will also greatly assist the study of archaeological remains, helping determine the place and time in which the domestication of llamas and alpacas would have occurred. With these results, it will be possible using a panel of 11 substitutions to distinguish a sample of a guanaco or vicuña furtively hunted, or a bone from a llama or alpaca in an archaeological site where its inhabitants have already moved from hunting wild camelids to the breeding of llamas and alpacas. Nevertheless, the absence of a solid association between the substitutions detected here and the different colours present in llamas and alpacas may be due to the existence of the participation of other genes in the expression of this trait. Future studies are needed to combine the analyses done here with similar assessments of polymorphisms in the α-Melanocyte-stimulating hormone (∝-MSH), the Tyrosinase-related protein (TRP1 and 2), the Membrane-Associated Transporter Protein (MATP), and the receptor tyrosine kinases (KIT) in the four species. Genomic studies, on the other hand, will also be able to shed more light on the network of genes involved in this particular phenotype.

ETHICS STATEMENT

Samples were collected following guidelines of the American Society of Mammalogists (Sikes et al., 2011). Specific Permits were required for the Servicio Agrícola y Ganadero, SAG (Permit 447, 2002), the Corporación Nacional Forestal, CONAF (Permit 6/02/2002), for granting other collection permits and helping in collecting samples. The animal research oversight committee of Universidad del Bío-Bío had knowledge of sampling plans prior to their approval of the present animal research protocol. All experimental protocols were approved by the Institutional Animal Care and Use Committee of Universidad del Bío-Bío, the methods were carried out in accordance with the approved guidelines.

AUTHOR CONTRIBUTIONS

JM developed the ideas and obtained funding for the project. RR, VV, and JM collected the samples. RR, JC, and AlC conducted the DNA analyses. JM, VV, AA, AnC, WJ, and POTW analysed the data. JM and WJ wrote the paper. All authors read, commented on and approved the final version of the manuscript.

FUNDING

This research was supported by FONDECYT, Chile Grant 1140785, Postdoctoral Grant 3050046 and CONICYT Chile (Beca de Apoyo a Tesis Doctoral), Morris Animal Foundation (D05LA-002), Darwin Initiative for the Survival of Species (United Kingdom) 1312 grant 162/06/126, The British Embassy (Lima), NERC 1313 (United Kingdom) grant GST/02/828, and Newton Fund Researcher Links Travel grant (ID: RLTG9-LATAM-359537872) funded by the UK Department for Business, Energy and Industrial Strategy and CONCYTEC (Peru) and delivered by the British Council.

ACKNOWLEDGMENTS

In Chile, we thank the Servicio Agrícola y Ganadero, SAG, the Corporación Nacional Forestal, CONAF for granting other collection permits. We also thank Jane C. Wheeler (CONOPA, Perú), Benito González (Universidad de Chile), Cristian Bonacic (Pontificia Universidad Católica de Chile), Pablo Valdecantos (Universidad Nacional de Tucumán), Alejandra von Baer (Llamas del Sur), Luis Jacome (Zoológico de Buenos Aires, Argentina), Alberto Duarte (Zoológico de Mendoza, Argentina), and Virginia Burgi and Ricardo Baldi (CEMPAT) for sharing samples. Special thanks to Kylie Ann Munyard (Curtin University of Technology), and Michael Bruford (Cardiff University) for useful information, discussions, and support. Samples were transported under CITES authorization numbers 6282, 4222, 6007, 5971, 5177, 5178, 23355, 22967, and 22920.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00487/full#supplementary-material

REFERENCES

Adeola, A. C., Oluwole, O. O., Oladele, B. M., Olorungbounmi, T. O., Boladuro, B., Olaogun, S. C., et al. (2017). Analysis of the genetic variation in mitochondrial DNA, Y-chromosome sequences, and MC1R sheds light on the ancestry of Nigerian indigenous pigs. *Genet. Sel. Evol.* 49:52. doi: 10.1186/s12711-017-0326-1

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., et al. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* 9:813. doi: 10.1038/s41467-018-03206-y

Almathen, F., Elbir, H., Bahbahani, H., Mwacharo, J., and Hanotte, O. (2018). Polymorphisms in *MC1R* and *ASIP* genes are associated with coat color variation in the arabian camel. *J. Hered.* 109, 700–706. doi: 10.1093/jhered/esy024

Bathrachalam, C., LaManna, V. L., Renieri, C., and LaTerza, A. (2011). "ASIP and MC1R cDNA polymorphism in alpaca," in *Fibre Production in South American Camelids and Other Fibre Animals*, eds M. Á. Pérez, J. P. Gutiérrez, I. Cervantes, and M. J. Alcalde (Amsterdam: Wageningen Academic Publishers), 93–96. doi: 10.3920/978-90-8686-727-1_11

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Carpio, M., and Solari, Z. (1982). Diámetro de la fibra en el vellón de la vicuña. *Informes de Trabajos de Investigación en Vicuñas UNALM Perú* 1, 54–102.

Cartajena, I., Núñez, L., and Grosjean, M. (2007). Camelid domestication on the western slope of the Puna de Atacama, northern Chile. *Anthropozoologica* 42, 155–173.

Chessa, B., Pereira, F., Arnaud, F., Amorim, A., Goyache, F., Mainland, I., et al. (2009). Revealing the history of sheep domestication using retrovirus integrations. *Science* 324, 532–536. doi: 10.1126/science.1170587

Cieslak, M., Reissmann, M., Hofreiter, M., and Ludwing, A. (2011). Colours of domestication. *Biol. Rev.* 86, 885–899. doi: 10.1111/j.1469-185x.2011.00177.x

Currat, M., Ruedi, M., Petit, R. J., and Excoffier, L. (2008). The hidden side of invasions: massive introgression by local genes. *Evolution* 62, 1908–1920. doi: 10.1111/j.1558-5646.2008.00413.x

Daveiro, M. S., Rigalt, F., Romero, S., Vidal, L., and DiRocco, F. (2016). Polymorphisms in MC1R and ASIP genes and their association with coat color phenotypes in llamas (*Lama glama*). *Small Rumin. Res.* 144, 83–89. doi: 10.1016/j.smallrumres.2016.08.003

Dong, Y., Zhang, X., Xie, M., Arefnezhad, B., Wang, Z., Wang, W., et al. (2015). Reference genome of wild goat (*Capra aegagrus*) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC Genom.* 16:431. doi: 10.1186/s12864-015-1606-1

Eizirik, E., Yuhki, N., Johnson, W. E., Menotti-Raymond, M., Hannah, S. S., and O'Brien, S. J. (2003). Molecular genetics and evolution of melanism in the cat family. *Curr. Biol.* 13, 448–453. doi: 10.1016/S0960-9822(03)00128-3

Fang, M., Larson, G., Ribeiro, H. S., Li, N., and Andersson, L. (2009). Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet.* 5:e1000341. doi: 10.1371/journal.pgen.1000341

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929–941. doi: 10.1534/genetics.112.147231

Feeley, N. L., Bottomley, S., and Munyard, K. A. (2011). Three novel mutations in ASIP associated with black fibre in alpacas (*Vicugna pacos*). *J. Agric. Sci.* 149, 529–538. doi: 10.1017/S0021859610001231

Feeley, N. L., and Munyard, K. A. (2009). Characterisation of the melanocortin-1 receptor gene in alpaca and identification of possible markers associated with phenotypic variations in colour. *Anim. Prod. Sci.* 49, 675–681. doi: 10.1071/AN09005

Foll, M., and Gaggiotti, O. E. (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Fontanesi, L., Beretti, F., Riggio, V., Dall'Olio, S., González, E. G., Finocchiaro, R., et al. (2009). Missense and nonsense mutations in melanocortin 1 receptor (MC1R) gene of different goat breeds: association with red and black coat colour phenotypes but with unexpected evidences. *BMC Genetics* 10:47. doi: 10.1186/1471-2156-10-47

Fontanesi, L., Dall'Olio, S., Beretti, F., Portolano, B., and Russo, V. (2011). Coat colours in the Massese sheep breed are associated with mutations in the agouti signaling protein (ASIP) and melanocortin 1 receptor (MC1R) genes. *Animal* 5, 8–17. doi: 10.1017/S1751731110001382

Franklin, W. L. (1982). "Biology, ecology and relationship to man of the South American camelids," in *Mammalian Biology in South America*, eds M. A. Mares and H. H. Genoways (Pittsburgh, PA: Pymatuning Laboratory of Ecology, PITT), 457–489.

Gepts, P., and Papa, R. (2002). "Evolution during domestication," in *Encyclopedia of Life Sciences*, ed. Macmillan Publishers Ltd. (London: Nature Publishing Group). doi: 10.1038/npg.els.0003071

Goldberg, A., Mychajliw, A. M., and Hadly, E. A. (2016). Post-invasion demography of prehistoric humans in South America. *Nature* 532, 232–235. doi: 10.1038/nature17176

Gratten, J., Beraldi, D., Lowder, B. V., McRae, A. F., Visscher, P. M., Pemberton, J. M., et al. (2007). Compelling evidence that a single nucleotide substitution in TYRP1 is responsible for coat-colour polymorphism in a free-living population of Soay sheep. *Proc Roy. Soc.* 274, 619–626. doi: 10.1098/rspb.2006.3762

Hemmer, H. (1990). *Domestication: The Decline of Environmental Appreciation*. Cambridge: Cambridge University Press.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Kadwell, M., Fernandez, M., Stanly, H. F., Baldi, R., Wheeler, J. C., Rosadio, R., et al. (2001). Genetic analysis reveals the wild ancestors of the llama and the alpaca. *Proc. Biol. Sci.* 268, 2575–2584. doi: 10.1098/rspb.2001.1774

Kerns, J. A., Newton, J., Berryere, T. G., Rubin, E. M., Cheng, J. F., Schmutz, S. M., et al. (2004). Characterization of the dog Agouti gene and a nonagouti mutation in German Shepherd dogs. *Mamm. Genome* 15, 798–808. doi: 10.1007/s00335-004-2377-1

Kijas, J. M., Wales, R., Törnsten, A., Chardon, P., Moller, M., and Andersson, L. (1998). Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics* 150, 1177–1185.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto-Neto, L. R., San-Cristol, M., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10:e1001258. doi: 10.1371/journal.pbio.1001258

Larson, G., and Burger, J. (2013). A population genetics view of animal domestication. *Trends Genet.* 29, 197–205. doi: 10.1016/j.tig.2013.01.003

Li, M.-H., Tiirikka, T., and Kantaden, J. (2014). A genome-wide scan study identifies a single nucleotide substitution in ASIP associated with white versus non-white coat-colour variation in sheep (*Ovis aries*). *Heredity* 112, 122–131. doi: 10.1038/hdy.2013.83

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

López, P., Cartajena, I., Santander, B., Rivera, B., and Opazo, C. (2012). Explotación de camélidos de un sitio intermedio tardío (1.000-1.400 d.C.) y tardío (1.400-1.536 d.C.) del Valle de Mauro (IV Región, Chile). *Boletín de la Sociedad Chilena de Arqueología* 4, 91–108.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A., Castaños, P., et al. (2009). Coat color variation at the beginning of horse domestication. *Science* 324:485. doi: 10.1126/science.1172750

Marín, J. C., Casey, C. S., Kadwell, M., Yaya, K., Hoces, D., Olazabal, et al. (2007a). Mitochondrial phylogeography and demographic history of the vicuña: implications for conservation. *Heredity* 99, 70–80. doi: 10.1038/sj.hdy.6800966

Marín, J. C., Romero, K., Rivera, R., Johnson, W. E., and Gonzáles, B. A. (2017). Y-chromosome and mtDNA variation confirms independent domestications and directional hybridization in South American camelids. *Anim. Genet.* 48, 591–595. doi: 10.1111/age.12570

Marín, J. C., Spotorno, A. E., Gonzáles, B. A., Bonacic, C., Wheeler, J. C., Casey, C. S., et al. (2008). Mitochondrial DNA variation and systematics of the guanaco (*Lama guanicoe*, Artiodactyla: Camelidae. *J. Mammal.* 89, 269–281. doi: 10.1644/06-MAMM-A-385R.1

Marín, J. C., Varas, V., Vila, A. R., López, R., Orozco-terWengel, P., and Corti, P. (2013). Refugia in patagonian fjords and the eastern Andes during the last glacial maximum revealed by huemul (*Hippocamelus bisulcus*) phylogeographical patterns and genetic diversity. *J. Biogeogr.* 40, 2285–2298. doi: 10.1111/jbi.12161

Marín, J. C., Zapata, B., Gonzáles, B., Bonacic, C., Wheeler, J. C., Casey, C., et al. (2007b). Systematics, taxonomy and domestication of alpaca and llama: new chromosomal and molecular evidence. *Rev. Chil. Hist. Nat.* 80, 121–140. doi: 10.1111/age.12570

Marklund, L., Moller, M. J., Sandberg, K., and Andersson, L. (1996). A missense mutation in the gene for melanocyte-stimulating hormone receptor (MC1R) is associated with the chestnut coat color in horses. *Mamm. Genome* 7, 895–899. doi: 10.1007/s003359900264

Norris, B. J., and Whan, V. A. (2008). A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Res.* 18, 1282–1293. doi: 10.1101/gr.072090.107

Powell, A. J., Moss, M. J., Tegland-Tree, L., Roeder, B. L., Carleton, C. L., Campbell, E., et al. (2008). Characterization of the effect of Melanocortin 1 Receptor, a member of the hair color genetic locus, in alpaca (*Lama pacos*) fleece color differentiation. *Small Rumin. Res.* 79, 183–187. doi: 10.1016/j.smallrumres.2008.07.025

Rieder, S., Taourit, S., Mariat, D., Langlois, B., and Guérin, G. (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (*Equus caballus*). *Mamm. Genome* 12, 450–455. doi: 10.1007/s003350020017

Royo, L. J., Alvarez, I., Arranz, J. J., Fernández, I., Rodríguez, A., Pérez-Pardal, L., et al. (2008). Differences in expression of the ASIP gene are involved in the recessive black coat colour pattern in sheep. Evidence from the rare Xalda sheep breed. *Anim. Genet.* 39, 290–293. doi: 10.1111/j.1365-2052.2008.01712.x

Russo, I. R. M., Hoban, S., Bloomer, P., Kotzé, A., Segelbacher, G., Rushworth, I., et al. (2018). 'Intentional genetic manipulation' as a conservation threat. *Conserv. Genet. Resour.* 1–11. doi: 10.1007/s12686-018-0983-6

Sambrook, J., Fritschi, E. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. New York, NY: Cold Spring Harbor Laboratory Press.

Schmutz, S. M., Berryere, T. G., Ellinwood, N. M., Kerns, J. A., and Barsh, G. S. (2003). MC1R studies in dogs with melanistic mask or brindle patterns. *J. Hered.* 94, 69–73. doi: 10.1093/jhered/esg014

Sikes, R. S., Gannon, W. L., and The Animal Care and Use Committee of the American Society of Mammalogists (2011). Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *J. Mammal.* 92, 235–253. doi: 10.1644/10-MAMM-F-355.1

Sinnwell, J. P., Schaid, D. J., and Yu, Z. (2007). *haplo. stats: Statistical Analysis of Haplotypes with Traits and Covariates When Linkage Phase is Ambiguous*. Available at: https://rdrr.io/cran/haplo.stats/

Stanley, H. F., Kadwell, M., and Wheeler, J. C. (1994). Molecular evolution of the family Camelidae: a mitochondrial DNA study. *Proc. R. Soc. Lond. B* 256, 1–6. doi: 10.1098/rspb.1994.0041

Våge, D. I., Klungland, H., Lu, D., and Cone, R. D. (1999). Molecular and pharmacological characterization of dominant black coat color in sheep. *Mamm. Genome* 10, 39–43. doi: 10.1007/s003359900939

Wang, G. D., Cheng, L., Fan, R., Irwin, D. M., Tang, S., Peng, J., et al. (2013). Signature of balancing selection at the MC1R gene in Kunming dog populations. *PLoS One* 8:e55469. doi: 10.1371/journal.pone.0055469

Wheeler, C. J. (2012). South American camelids - past, present and future. *J. Camelid Sci.* 5, 1–24. doi: 10.3389/fpls.2018.00649

Wheeler, J. C. (1991). "Origen, evolución y status actual," in *Avances y perspectivas del conocimiento de los camélidos Sudamericanos*, ed. S. Fernandez (Santiago: FAO), 11–48.

Wheeler, J. C. (1995). Evolution and present situation of the South American camelidae. *Biol. J. Linn. Soc.* 54, 271–295. doi: 10.1016/0024-4066(95)90021-7

Wheeler, J. C. (2004). Evolution and present situation of the South American camelidae. *Biol. J. Linn. Soc.* 54, 271–295. doi: 10.1016/0024-4066(95)90021-7

Wheeler, J. C., Russel, A. J., and Stanley, H. F. (1992). A measure of loss: prehispanic llama and alpaca breeds. *Archivos de Zootecnia* 41:17.

Wheeler, J. C., Russel, A. J. F., and Redden, H. (1995). Llamas and alpacas: pre-conquest breeds and post-conquest hybrids. *J. Archaeol. Sci.* 22, 833–840. doi: 10.1016/0305-4403(95)90012-8

Wu, X., Tan, Z., Shen, L., Yang, Q., Cheng, X., Liao, K., et al. (2017). Coat colour phenotype of Qingyu pig is associated with polymorphisms of melanocortin receptor 1 gene. *Asian Australas. J. Anim. Sci.* 30, 938–943. doi: 10.5713/ajas.16.0376

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor is currently co-organising a Research Topic with one of the authors POTW, and confirms the absence of any other collaboration.

Copyright © 2018 Marín, Rivera, Varas, Cortés, Agapito, Chero, Chávez, Johnson and Orozco-terWengel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 05 February 2019
doi: 10.3389/fgene.2019.00048

Cdrom Archive: A Gateway to Study Camel Phenotypes

Hasan Alhaddad and Bader H. Alhajeri*

Department of Biological Sciences, Kuwait University, Kuwait City, Kuwait

Camels are livestock that exhibit unique morphological, biochemical, and behavioral traits, which arose by natural and artificial selection. Investigating the molecular basis of camel traits has been limited by: (1) the absence of a comprehensive record of morphological trait variation (e.g., diseases) and the associated mode of inheritance, (2) the lack of extended pedigrees of specific trait(s), and (3) the long reproductive cycle of the camel, which makes the cost of establishing and maintaining a breeding colony (i.e., monitoring crosses) prohibitively high. Overcoming these challenges requires (1) detailed documentation of phenotypes/genetic diseases and their likely mode of inheritance (and collection of related DNA samples), (2) conducting association studies to identify phenotypes/genetic diseases causing genetic variants (instead of classical linkage analysis, which requires extended pedigrees), and (3) validating likely causative variants by screening a large number of camel samples from different populations. We attempt to address these issues by establishing a systematic way of collecting camel DNA samples, and associated phenotypic information, which we call the "Cdrom Archive." Here, we outline the process of building this archive to introduce it to other camel researchers (as an example). Additionally, we discuss the use of this archive to study the phenotypic traits of Arabian Peninsula camel breeds (the "Mezayen" camels). Using the Cdrom Archive, we report variable phenotypic traits related to the coat (color, length, and texture), ear and tail lengths, along with other morphological measurements.

Keywords: camel biobank, camel breed, camel ear, coat color, hair length, hair texture, Mezayen, tail length

OPEN ACCESS

Edited by:
Pamela Burger,
University of Veterinary Medicine
Vienna, Austria

Reviewed by:
Pablo Orozco-terWengel,
Cardiff University, United Kingdom
Margarida Matos,
Universidade de Lisboa, Portugal

****Correspondence:***
Hasan Alhaddad
hassan.alhaddad@ku.edu.kw

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *15 October 2018*
Accepted: *21 January 2019*
Published: *05 February 2019*

Citation:
Alhaddad H and Alhajeri BH (2019) Cdrom Archive: A Gateway to Study Camel Phenotypes. Front. Genet. 10:48. doi: 10.3389/fgene.2019.00048

INTRODUCTION

Dromedary camels (*Camelus dromedarius* Linnaeus, 1758) are exceptional livestock animals because of their natural adaptations to hot sandy desert environments (Schmidt-Nielsen, 1959; Abu-seida et al., 2012; Eshra and Badawy, 2014) and their artificially selected traits (Farah, 1993; Khalaf, 1999; Kadim et al., 2008; Teague, 2009; Kagunyu et al., 2013).

Despite the seemingly large variation in physiological, biochemical, morphological, and behavioral traits, the camel has received little attention with regard to the documentation of these traits, insofar as their hereditary status and their molecular basis (Burger, 2016). Using various genetic resources (Al-Swailem et al., 2010; Wu et al., 2014; Fitak et al., 2016), few studies have recently started to investigate the genetic basis of camel phenotypic and behavioral traits (Holl et al., 2017; Almathen et al., 2018; Ramadan et al., 2018); mostly using the candidate gene(s) sequencing approach (Zhu and Zhao, 2007).

For example, sequencing the *KIT* (Tyrosine kinase receptor) gene revealed the variants associated with the white-spotting phenotype of piebald (painted) camels (Holl et al., 2017; Volpato et al., 2017). The candidacy of this gene was established based on findings in other animals. The *KIT* gene has been identified or implicated to be related to white color or white-spotting in alpacas (Jackling et al., 2014), cows (Fontanesi et al., 2010), yaks (Zhang et al., 2014), pigs (Cho et al., 2011), goats (Nazari-Ghadikolaei et al., 2018), horses (Hauswirth et al., 2013), donkeys (Haase et al., 2015), cats (David et al., 2014), dogs (Wong et al., 2012), mice (Geissler et al., 1988), and rabbits (Fontanesi et al., 2014). Thus, applying a candidate gene approach to camels requires the presence of the phenotype in other mammals and a manageable number of candidate genes to be sequenced.

Beyond the candidate gene approach (which requires the existence of a similar phenotype in a related mammal), genetic investigations in camels also includes classical linkage analysis (Ott et al., 2015), genome-wide association (Hirschhorn and Daly, 2005), or whole-genome sequencing approaches (Petersen et al., 2017). All these approaches provide an opportunity to study camel-specific characteristics, and for many cases, narrow down the number of candidate genes to investigate. However, several challenges hinder the implementation of these approaches in camels. These challenges include: (1) the limited camel genetic resources (i.e., no high-density SNP array or genome-wide STR panel), (2) the lack of multigenerational pedigrees to conduct linkage analyses, (3) the difficulty of obtaining a pedigree when most camel breeders rely on mental documentation of their crosses (Köhler-Rollefson, 1993), (4) the late breeding age (∼4 years) and the long gestation (∼12 months) and weaning (∼9 months) periods of camels (which prevents any attempt to start a large scale breeding experiment) (Ali et al., 2018), (5) the absence of a detailed record of camel traits or genetic diseases and their likely mode of inheritance (i.e., dominant, recessive, etc.) and heritability, and (6) the lack of camel breed registry or recorded information, especially for desired traits (i.e., milk volume, meat quality, coat color, racing performance, etc.).

All the aforementioned genetic approaches to study camel phenotypes, as well as validation of phenotype-genotype association, require a large number of carefully phenotyped individuals of known ancestry. This necessity justifies the assembly of a camel DNA biobank, which is implemented in other livestock animals (Groeneveld et al., 2016; Blackburn, 2018). Accordingly, we established such a biobank, which we refer to as the *C. dromedarius* Archive ("Cdrom Archive") that consists of biological specimens (DNA source) accompanied by detailed specimen-associated information, such as age, sex, breed/type, pedigree, location, and a comprehensive documentation of morphological phenotypes in the form of photographs.

In this review, we present our methodology of collecting and organizing each camel sample in the archive. We also use the current samples of the Cdrom Archive to characterize six camel breeds from the Arabian Peninsula (Majaheem, Sofor, Shaele, Homor, Shageh, and Waddeh), with an emphasis on the variation in the coat (i.e., color, length, and texture), ear morphology (i.e., shape and length), and tail length.

BUILDING AND USING THE CDROM ARCHIVE

Data Collection and Organization

Sample-specific information of the Cdrom Archive is collected and organized in a unified format using the *SamplEase* application (Alhaddad and Alhajeri, 2018). While we propose to collect and organize our camel specimens using the aforementioned sample collection application, data can also be included in the Cdrom Archive manually. The archive is currently comprised of 163 samples that were collected during 2015 (February–April), 2016 (October–December), and 2017 (March–April) (**Figure 1A** and **Supplementary Table S4**). We plan to continue to add more samples (and associated phenotypic data) to the Cdrom Archive in the future. Our long-term plan is to make the archive available in a database on the web, which will continuously be updated with new specimens, as they are collected. The current Cdrom Archive specimen's information is listed in **Supplementary Table S4**, both photographs and biological material associated with each specimen is available upon request.

Sex Information

Females ($n = 131$) currently represent the majority of the samples of the Cdrom Archive. The discrepancy in the number of female to male ($n = 19$) samples is a consequence of each breeder keeping only one or two reproductively active males in their stock at each time (Ali et al., 2018) (**Figure 1A**).

Age Information

Samples from camels of various ages were collected. However, the majority of the samples thus far were of unknown age ($n = 83$) (**Figure 1B**). This is in part due to the lack of a written record of the breeding stock, and the reliance on teeth appearance and camel behavior to determine age. Most camel breeders in the Arabian Peninsula do not keep track of the specific age of each of their camels (i.e., number of years), but rather label their age class generally based on their behavior, reproductive maturity, and teeth development (**Figure 2** and **Supplementary Table S1**). It is thus necessary to use age categories such as juvenile, subadult, and adult instead of years. It is always possible to deduce the age category of each camel sampled in the Cdrom Archive by referring to the associated photographs (see below).

Pedigree Information

Pedigree information of Cdrom samples is mostly incomplete—it currently contains only seven trios (parents and an offspring) and 21 half-siblings (siblings sharing a single parent) (**Figure 1C**). It is difficult to obtain pedigreed camel samples because breeders in the Arabian Peninsula (1) rely mostly on a mental record of their breeding programs (Köhler-Rollefson, 1993), (2) assign the same name to multiple camels (which increases the likelihood of pedigree mistakes), (3) constantly exchange/sell camels with other breeders, and (4) use reproductively superior bull camels for breeding (bulls are neither owned by the breeder nor present at the time of sample collection). It is thus easier to locate and

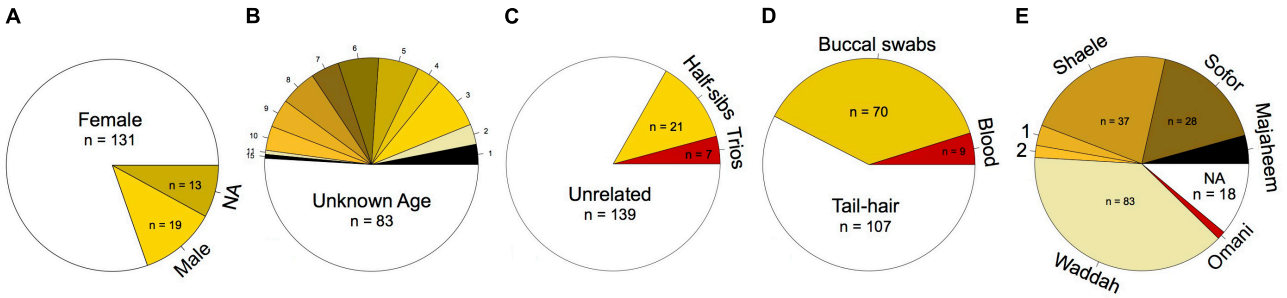


FIGURE 1 | Cdrom Archive sample-specific information. General counts based on **(A)** sex, **(B)** age, **(C)** pedigree information, **(D)** biological specimen type, and **(E)** breed affiliation (1: Homor and 2: Shageh).

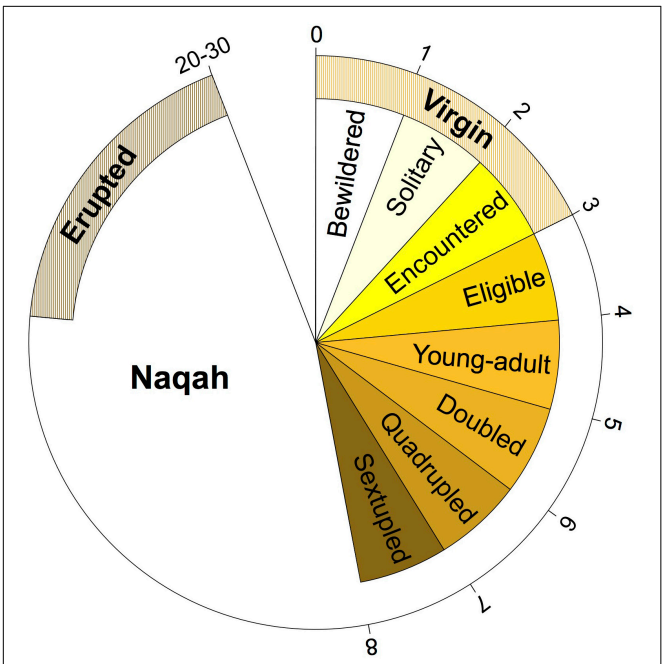


FIGURE 2 | Terminology of Mezayen camel age classes and their approximate age in years. Each age class represents a phase in the growth of a female camel (male terminology is similar) and does not depend on specific number of months/years. Outer numbers are approximate number of years for the completion of each age stage. For details regarding the terms and their meaning refer to **Supplementary Table 1**.

collect trios, siblings, half-siblings, or small pedigrees than find a multigenerational pedigree.

Biological Specimens

The biological specimens of the Cdrom Archive presently come from whole-blood, buccal swabs, and tail-hair (**Figure 1D**). We found that the most appropriate camel DNA source for the Cdrom Archive is tail-hair follicles—this is based on its ease of collection, transport, and storage, and because it provided adequate DNA quantities for genetic analyses 30 tail-hail follicles ≈6 μg (Alhaddad et al., 2019). The quantity of DNA, obtained from hair follicles, is thus expected to be successfully used in each of PCR, STR and SNP genotyping, targeted sequencing, and whole-genome sequencing.

In the process of establishing the Cdrom Archive, we arrived at the following recommendations to safely collect tail-hair samples (intended as a DNA source). To avoid startling the camel, it should be approached slowly from the front, and then it is advisable to pet the animal to allow it to relax, before moving toward the tail to collect the DNA sample. It was easier to collect tail-hair samples from females, since they tend to be more relaxed, probably since they are used to being milked by the breeders. Unlike horses that kick posteriorly, camels kick sideways, and thus, it is advisable when collecting tail-hair samples to stand behind the camel, and not to its side. To collect hair in an optimal manner, a small bundle of long tail-hairs near the base of the tail can be wrapped around the index finger and plucked upward. It is recommended to bind the hair bundle using tape and discard excess hair away from the roots (tips) (since it does not contain any DNA), before being stored in a labeled envelope.

Geographic Distribution

GIS coordinates are automatically assigned to each collected specimen in the Cdrom Archive using *SamplEase* (Alhaddad and Alhajeri, 2018). Most of our samples so far were collected from Kuwait (15 locations) and only nine samples come from Saudi Arabia (all from Alhasa) **Figure 3** – map generated using *ggmap* R package (Kahle and Wickham, 2013; R Development Core Team, 2018). We acknowledge that camel herds are generally maintained in an open environment, rather than in a closed farm, and that camel breeders change their location several times to prevent disease due to accumulation of fecal material and to void depleting grazing grounds. Nonetheless, GPS coordinates can be used to accurately reference each sample to its location of collection—this data may allow for the construction of a camel locality heat map, that would be helpful for national census surveys of camel populations, along with disease management and prevention plans e.g., managing the Middle East respiratory syndrome-MERS (Omrani et al., 2015).

Photographs

SamplEase allows for the collection of an unlimited number of photographs for each sampled camel, which are all linked to the

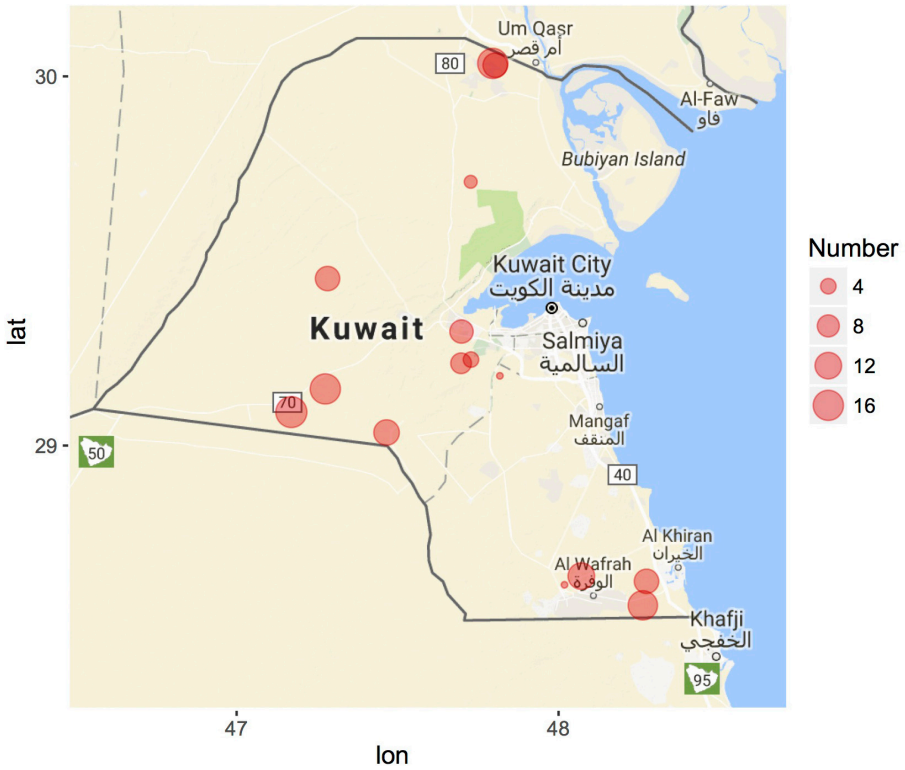


FIGURE 3 | Geographic distribution of Cdrom Archive samples from Kuwait. Samples were collected during 2015 (February–April), 2016 (October–December), and 2017 (March–April). Circle size corresponds to the number of camel samples from each breeder (location). Nine samples were collected from King Faisal University, Saudi Arabia (not shown). Map generated using the *ggmap* R package.

FIGURE 4 | Ear length and shape variation between Mezayen camel breeds. **(A)** Majaheem camels have a distinct long-pointed ear shape, referred to as "speared" ears, whereas **(B)** Malaween breeds (Sofor, Shaele, Homor, Shageh, and Waddah) all exhibit shorter ears that are "folded" or "tilted" sideways and to the back. The white and black disks are five centimeters in diameter, which were added to extract a scale factor in subsequent morphometric analyses. Images were extracted from Cdrom Archive photos (collected by the authors).

basic information for each camel sample. The majority of the sampled camels in the Cdrom Archive have been photographed multiple times—these photographs allow us to subsequently characterize the morphological features of each sampled camel.

Sampled Breeds in the Cdrom Archive

Most of our samples presently come from Kuwait and consist of camel breeds common in the Arabian Peninsula. The breeds currently in the Cdrom Archive are Majaheem, Sofor, Shaele, Homor, Shageh, Waddah, and Omani (**Figure 1E**). Many alternative spellings for these breeds exist in the literature; for consistency, we have adopted the spellings used by Porter et al. (2016).

Studying the molecular basis of any trait is more achievable in a breed rather than in a random bred population (Karlsson and Lindblad-Toh, 2008). This is due to the genetic similarity

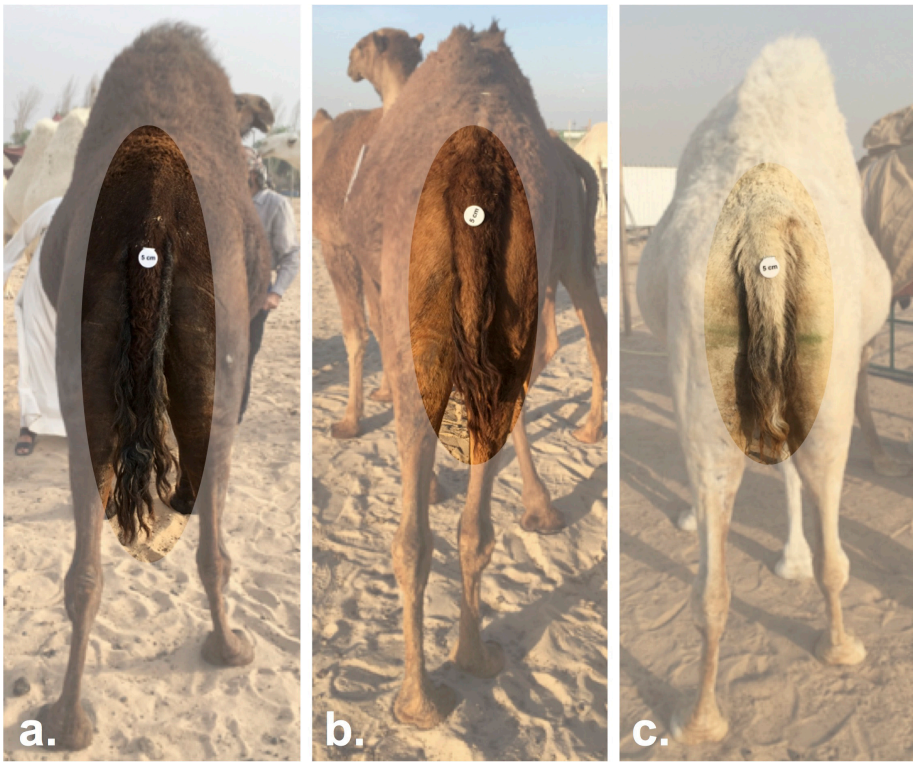


FIGURE 5 | Tail length variation between Mezayen camel breeds. **(a)** Majaheem camels are characterized by long tails with a narrow tail-base. **(b)** Shaele, and **(c)** Waddah camels, which represent the Malaween breeds, display short-tails with wide tail-bases. The white disks on the tail are reference scales (five centimeters in diameter). Images were extracted from Cdrom Archive photos (collected by the authors).

between individuals within a breed compared to an admixed population. The genetic similarity between members of a breed reduces the variable sites to be investigated and enables better localization of phenotype-associated genes. However, the concept of a breed is a subject of historic (Lloyd-Jones, 1915) and ongoing debates (Food Agriculture Organization of the United Nations, 2013), and applying this concept to dromedary camels is even more debatable and harder to implement (Köhler-Rollefson, 1993; Wardeh, 2004; Dioli, 2016). Animal breeds are generally defined based on characteristics agreed upon by breeders that are implemented using documented breed standards, which requires an animal registry, and a governing breed association (Food Agriculture Organization of the United Nations, 2013). The camel breeding community suffers from the lack any breeders' associations or organizations – such communities often set breed defining criteria and features for other animals. The closest to a camel breed registry or a governing body is the Camel Race Federation in the United Arab Emirates (Khalaf, 1999). However, the federation is mainly focused on racing camels, and is specialized in implementing rules for fair racing, rather than defining breed standards.

The closest to "true" Arabian Peninsula camel breeds are the "Mezayen" camels, a term that literally means "beauty-contest" camels. The Mezayen camel breeds are the: Majaheem, Sofor, Shaele, Homor, Shageh, and Waddah (Abdallah and Faye, 2012; Porter et al., 2016; Alaibil Festival, 2017). We argue for their breed status because (1) each breed is defined by a distinct color group and a set recognized morphological features (Köhler-Rollefson, 1993), (2) a consensus of breed standards exists among breeders specifically for these six breeds (Teague, 2009), and because (3) an incentive to maintain breed standards is available in the form of camel beauty and breeding excellence competitions, such as the highly prized camel beauty competition of the King Abdulaziz Camel Festival (Alaibil Festival, 2017), along with more regional/tribal competitions (Hammond, 2007).

MEZAYEN PHENOTYPES

The phenotypes and breed designations of domesticated animals are often more easily recognized by the breeder who selected for the particular traits. As such, we sought out Mezayen camel breeders to help in identifying and explaining the phenotypes of their camels that have been targets of selection using their common terminology. Mezayen camel breeders in the Arabian Peninsula use specific names to describe each breed (see above), breed subtype, and external phenotypes (**Supplementary Tables S2**, **S3**). The breed names and phenotypes described here are based on translations of the breeders' Arabic terminology to ensure correct breed and phenotype assignments when collecting Cdrom Archive samples (see **Supplementary Tables S2**, **S3** for details).

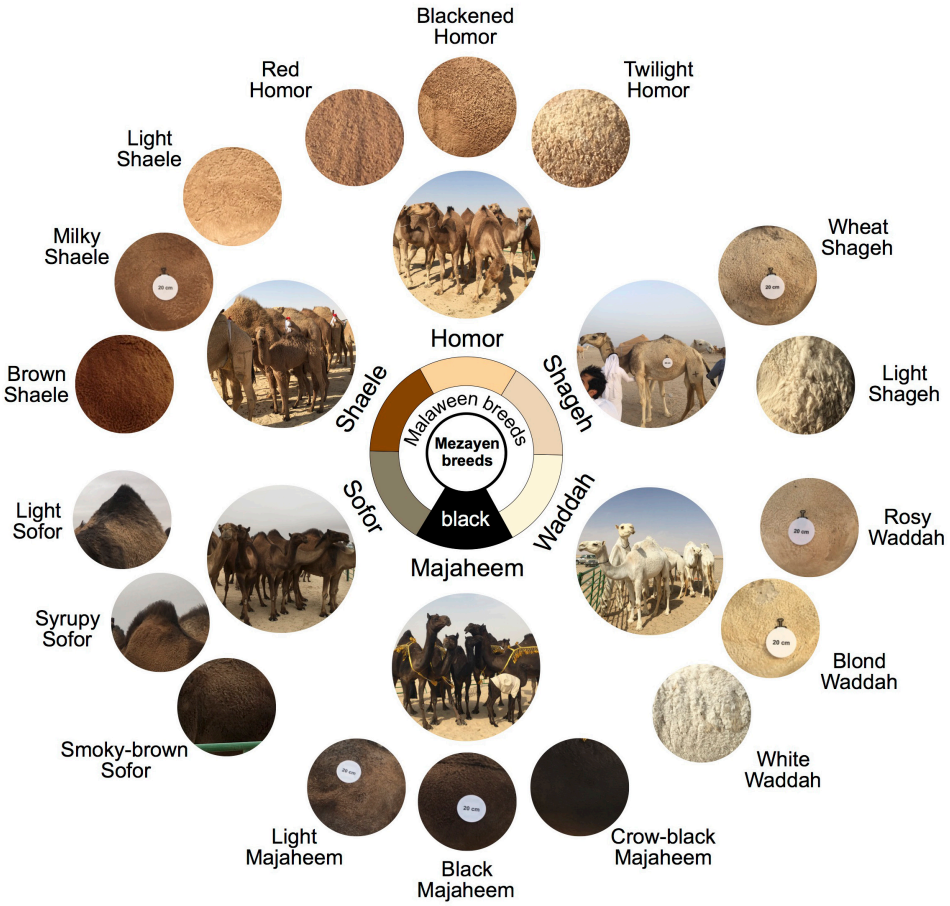


FIGURE 6 | Mezayen camel breeds of the Arabia Peninsula and their coat colors. The six breeds are primarily divided into black vs. Malaween breeds. This division is based on coat color and morphological characteristics (e.g., ear shape and tail length). The black breed is Majaheem whereas the Malaween camels are further divided into five breeds based their coat color. The Malaween breeds, ranging from dark to light, are as follows: Sofor, Shaele, Homor, Shageh, and Waddah. The outer (small) circles represent the "sub-colors" of each breed. The Majaheem sub-colors are crow-black, black, and light. The Sofor sub-colors are smoky-brown, syrupy, and light. The Shaele sub-colors are brown, milky, and light. The Homor sub-colors are red, blackened, and twilight. The Shageh sub-colors are wheat and light. The Waddah sub-colors are rosy, blond, and white. Coat color circles come from the Cdrom Archive photographs and were extracted from the part of the lateral torso that is below the hump. The white disks are reference scales (20 centimeters in diameter). Breed photos were taken by Hasan Alhaddad during Mutair Cultural Festival 2017, Kuwait. Coat color circles were extracted from Cdrom Archive photos (collected by the authors).

Mezayen camels are divided into two main groups, the dark colored Majaheem, and the "Malaween," which translates to colored breeds (Sofor, Shaele, Homor, Shageh, and Waddah) (personal observation). The separation of these two groups is in part based on coat color, but is also based on general features, such as body size, ear length and shape, and tail characteristics (Köhler-Rollefson, 1993; Abdallah and Faye, 2012). Majaheem camels are generally larger, and have long "speared" ears (**Figure 4A**), and a long tail with a narrow tail-base (**Figure 5a**) (Al-Hazmi et al., 1994). On the other hand, all Malaween breeds exhibit comparatively smaller body sizes, have short and tilted ears (**Figure 4B**), and a short tail with a wide tail-base (**Figures 5b–c**) (see **Supplementary Table S2** for naming details). Breeders often do not breed Majaheem camels with any of the Malaween breeds, and when such an event occurs, breeders can easily recognize the hybrid due to changes in body features; such hybrids are often disqualified from competing in beauty competitions (personal observation). The Malaween are subdivided based on their coat color (Porter et al., 2016).

Coat Color

Each Mezayen camel breed represents a color class (major color under which several varieties exist) (**Figure 6**). Broadly, the six color classes are black (Majaheem), smoky-brown (Sofor), brown (Shaele), red (Homor), wheat (Shageh), and white (Waddah) (Porter et al., 2016). Within each breed, a number of subtypes exist, which correspond to fine differences in coat color tone (**Figure 6** – outer circle) (see **Supplementary Table S2** for naming details). For example, under the broad black color class of the Majaheem, three subgroups are recognized. The sub-colors of Majaheem are (1) "crow-black" Majaheem, which as the name suggests, have a black coat color similar to the "blackness" of crow feathers, (2) "black" Majaheem are referred to by breeders as black, but is dark-brown color, that is

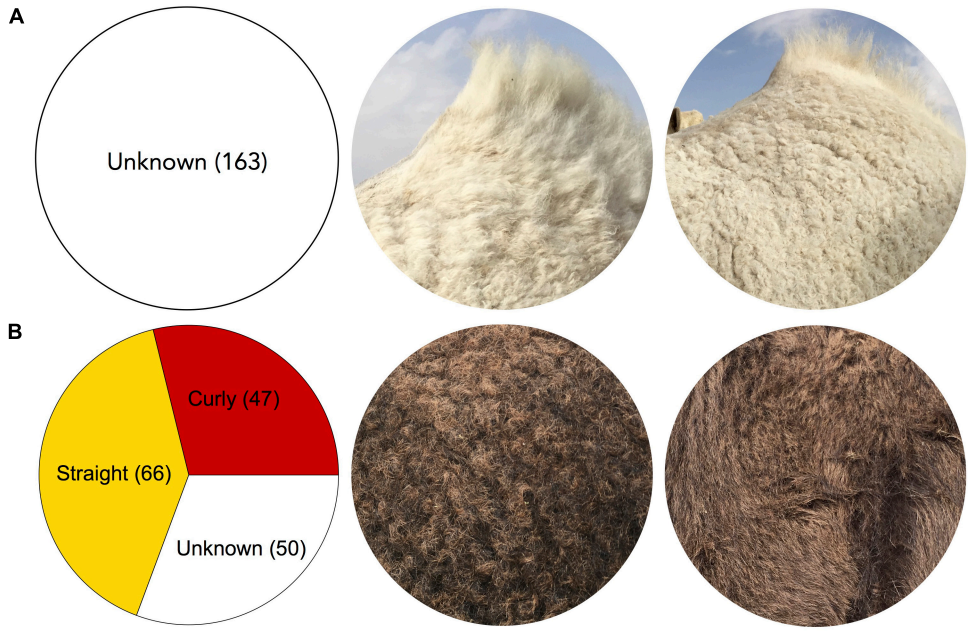


FIGURE 7 | Camel coat length and texture variation. **(A)** Long and short coats, respectively. **(B)** "Ringed" (curly) and straight camel coat textures, respectively. The Cdrom Archive contains samples of both identified coat texture phenotypes from various Mezayen breeds **(B)**. Images were extracted from Cdrom Archive photos (collected by the authors).

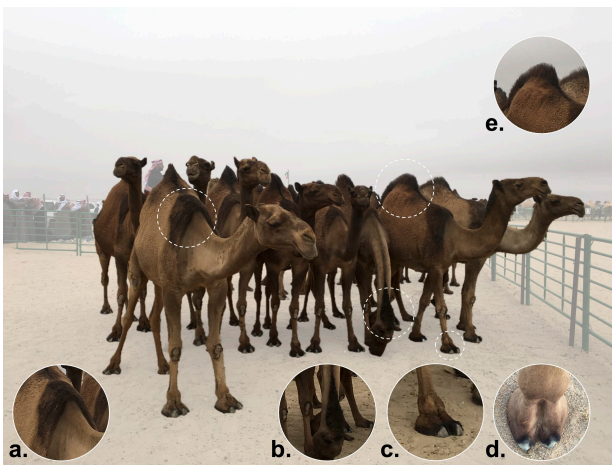


FIGURE 8 | Dark coat "highlights" of the "Syrupy" Sofor phenotype. "Syrupy" Sofor camels show a darker coat color, located on **(a)** the withers, **(b)** the top of the neck, **(c,d)** the dorsal footpad and toes, and **(e)** the tip of the hump. The name of this breed is inspired by the camel breeders' imagination, where the camel extremities "seem" as if they were "dipped in date syrup." Photo was taken by Hasan Alhaddad during Mutair Cultural Festival 2017, Kuwait.

similar to darkly roasted coffee beans, and (3) "light" Majaheem have a dark brown coat color with scattered light-colored hairs.

Camel coat colors have been recently investigated by sequencing two candidate genes *MC1R* (*melanocortin 1 receptor*) and *ASIP* (*agouti signaling protein*) (Almathen et al., 2018). Polymorphisms within the two candidate genes are found to be associated with broad color classifications (i.e., a single variant for black and dark brown colors) (Almathen et al., 2018). The color classifications presented here are more refined and are suspected to identify additional associated variants within *MC1R* and *ASIP* of each color (if they exist) or unravel a more complex genetic basis of coat color in camels.

Hair Length

Two hair length varieties (short and long) exist in each of the six Mezayen camels (**Figure 7A**). Breeders least favor the long-haired variety of each breed, especially when the hair texture is straight (personal observation). Thus, the identification of the molecular basis of hair length in camels may aid breeders in selecting camels to breed based on their genotype.

Fibroblast growth factor-5 (*FGF5*) has been identified to be associated with, or responsible for, long hair in many mammals, including llamas (Daverio et al., 2017), alpacas (Pallotti et al., 2018), donkeys (Legrand et al., 2014), sheep (Zhang et al., 2015), goats (Wang et al., 2016), cats (Drögemüller et al., 2007), dogs (Dierks et al., 2013), rabbits (Mulsant et al., 2010), mice (Hébert et al., 1994), and humans (Higgins et al., 2014). It is thus justifiable to consider *FGF5* as a strong candidate gene for long hair in Mezayen camels, and this hypothesis can be investigated via direct sequencing.

Hair Texture

The hair texture of Mezayen camel coats comes into two varieties, straight and ringed (**Figure 7B**). These two varieties occur in all six breeds. Breeders select for curly hair that appear as rings,

especially in the torso region, which are considered signs of beauty and health (personal observation). To achieve the most desirable coat for beauty competitions, breeders often select for a combination of short and ringed coat hairs (personal observation). The Crdom Archive currently contains 66 straight hair camels and 47 ringed hair camels, which we aim to use in genetic association studies—this relatively large sample size is optimal since a large number of genes are expected to be responsible for a curly coat (**Figure 7B**).

"Syrupy" Sofor Coat Color

The "Syrupy" Sofor displays a unique coat color phenotype (see **Supplementary Table S2** for naming details). This Sofor camel subtype shows a darker coat pigmentation at some body extremities, such as the withers, upper neck, dorsal footpad, nails, tip of the hump, and the tail (**Figure 8**). This phenotype does not occur in light colored breeds (Homor, Shagah, Waddah), but occasionally occurs in the Shaele breed, due to its intercrossing with the Sofor breed. This color phenotype has equivalents in other mammals, such as the "points" coat phenotype of Siamese and Burmese cats (Lyons et al., 2005), California rabbits (Aigner et al., 2000), and mice (Beermann et al., 2004).

Mutations in the *Tyrosinase* (*TRY*) gene have been associated with darker coloration in specific body parts, which arises due to the temperature sensitivity of gene production (Lyons et al., 2005). The close resemblance in coat phenotype between Syrupy Sofor camels, Siamese and Burmese cats, California rabbits, and mice, suggests that the *TYR* gene could be a strong candidate for this phenotype. Direct sequencing of the Syrupy Sofor camel genes, and the sequencing of the genes of their counterparts of the same breed (smoky-brown and light Sofor) could be a direct approach to study this phenotype.

CAMEL MORPHOMETRICS

Several studies have examined the variation in body measurements among camel breeds (Al-Hazmi et al., 1994; Abdallah and Faye, 2012). So far, most published studies that investigate this theme use traditional, distance-based approaches, using calipers and measuring tape. While including such data along with each Cdrom Archive sample would provide valuable insights into the extent of the morphometric differentiation among the breeds, based on our personal experience, collecting such data manually is time-intensive and imprecise, given the temperamentality of most of the camels that we handled. Consequently, we developed a standardized method of photographing the sampled camels using the *SamplEase* application, where photographs are taken in such a way as to allow for the extraction of both linear and geometric morphometric data. We attach a scale bar to each sampled camel prior to photography to allow for the extraction of a scaling factor, which allows for the conversion of pixels to real units (i.e., centimeters). The "geometric morphometric" approach of examining morphological variation is commonly employed to extract data from zoological specimens (Zelditch et al., 2004; Alhajeri, 2018), and has recently been used to characterize morphological variation in live horses (Druml et al., 2015). More advanced methods of quantifying morphometric variation in camels in three-dimensions may also be implemented in the future.

CONCLUSION

This review focused on outlining the framework of building and sample collection of our recently developed Cdrom Archive. This outline was intended to provide an example of how to establish a biobank that would be useful for genetic studies, thus we hope it would encourage others to establish similar camel biobanks elsewhere. Using the samples collected thus far, we introduced six camel breeds of the Arabian Peninsula that are used in camel beauty competitions and referred to as the Camel Mezayen contest. Using the photographs of the Cdrom archive, we discussed the coat color variations and their naming, as well as ear and tail variation. Where applicable, we outlined possible genetic approaches to study the genetics of these phenotypes and suggested likely candidate genes. Lastly, we introduced the possibility of applying morphometric tools to extract data from the photographs of the Cdrom Archive, which would allow us to investigate body size and shape variation. This review aimed to provide an example of what can be done across camel research laboratories to collect and characterize camel phenotypes, and possibly traits associated with production and adaptation for future genetic studies.

AUTHOR CONTRIBUTIONS

HA and BA collected the samples and wrote the manuscript.

FUNDING

No part of this work, including travel and sample collection, received support from grant funding.

ACKNOWLEDGMENTS

We are grateful to Mohsen Bin Fawaz Alshaibani, the camel expert and judge in Mezayen festivals, for his valuable information and clarifications of the breed types, subtypes, and phenotypes. We also thank the organizers of the Mutair cultural festival (2017) for granting us access to the camel shows and the judging arena. We are indebted to Kuwaiti camel breeders (Abdulatif Alhedari, Anwar Alfadhli, Mohammed Alefasi, Nawaf Alienizi, Adel Alothman, Waleed Alqallaf, Theyab Alhajeri, Abdullah Alefasi, Saad Alajmi, Mutlaq Alsegyani, Adnan Alshawaf, and Saree Alhajeri) and Faisal Almathen (King Faisal University) for providing camel samples and information.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00048/full#supplementary-material

TABLE S1 | Camel terminology used to describe female camel age classes. Age terms were translated from original Arabic terms and the Arabic pronunciation is shown in italics. Similar names are given to male camels with slight differences related to gender changes to original terms in Arabic.

TABLE S2 | Names of "Mezayen" camel breeds and their subtypes. Names of breeds and subtypes were translated from original Arabic terms and the Arabic pronunciation shown in italics. The translations are near exact to what is offered by the breeders, and in some cases the names do not exactly reflect the actual appearance. For visual comparison refer to **Figure 6**.

TABLE S3 | Terminology of Mezayen camel coat texture and ear shape. Phenotype terminologies were translated from original Arabic terms and the Arabic pronunciation shown in italics.

TABLE S4 | Current sample list and information of the Cdrom Archive.

REFERENCES

Abdallah, H. R., and Faye, B. (2012). Phenotypic classification of Saudi Arabian camel (*Camelus dromedarius*) by their body measurements. *Emirates J. Food Agric.* 24, 272–280.

Abu-seida, A., Mostafa, A., and Tolba, A. R. (2012). Anatomical and Ultrasonographical studies on tendons and digital cushions of normal phalangeal region in camels (*Camelus dromidarius*). *J. Camel Pract. Res.* 19, 169–175.

Aigner, B., Besenfelder, U., Muller, M., and Brem, G. (2000). Tyrosinase gene variants in different rabbit strains. *Mamm. Genome* 11, 700–702. doi: 10.1007/s003350010120

Alaibil Festival (2017). *King Abdulaziz Camel Festival*. Available: http://www.alaibilfestival.com/en/

Alhaddad, H., and Alhajeri, B. H. (2018). SamplEase: a simple application for collection and organization of biological specimen data in the field. *Ecol. Evol.* 8, 10266–10271. doi: 10.1002/ece3.4503

Alhaddad, H., Maraqa, T., Alabdulghafour, S., Alaskar, H., Alaqeely, R., Almathen, F., et al. (2019). Quality and quantity of dromedary camel DNA sampled from whole-blood, saliva, and tail-hair. *PLoS One* 14:e0211743. doi: 10.1371/journal.pone.0211743

Alhajeri, B. H. (2018). Cranial variation in geographically widespread dwarf gerbil *Gerbillus nanus* (Gerbillinae, Rodentia) populations: isolation by distance versus adaptation to local environments. *J. Zool. Syst. Evol. Res.* 9, 928–937. doi: 10.1111/jzs.12247

Al-Hazmi, M., Ghandour, A., and Elgohar, M. (1994). A study of the biometry of some breeds of arabian camel (came/us dromedarius) in Saudi Arabia. *J. King Abdulaziz Univ.* 6, 87–99. doi: 10.4197/Sci.6-1.7

Ali, A., Derar, D., Alsharari, A., Alsharari, A., Khalil, R., Almundarij, T. I., et al. (2018). Factors affecting reproductive performance in dromedary camel herds in Saudi Arabia. *Trop. Anim. Health Prod.* 50, 1155–1160. doi: 10.1007/s11250-018-1545-3

Almathen, F., Elbir, H., Bahbahani, H., Mwacharo, J., and Hanotte, O. (2018). Polymorphisms in MC1R and ASIP genes are associated with coat colour variation in the Arabian camel. *J. Hered.* 109, 700–706. doi: 10.1093/jhered/esy024

Al-Swailem, A. M., Shehata, M. M., Abu-Duhier, F. M., Al-Yamani, E. J., Al-Busadah, K. A., Al-Arawi, M. S., et al. (2010). Sequencing, analysis, and annotation of expressed sequence tags for *Camelus dromedarius*. *PLoS One* 5:e10720. doi: 10.1371/journal.pone.0010720

Beermann, F., Orlow, S. J., and Lamoreux, M. L. (2004). The Tyr (albino) locus of the laboratory mouse. *Mamm. Genome* 15, 749–758. doi: 10.1007/s00335-004-4002-8

Blackburn, H. D. (2018). Biobanking genetic material for agricultural animal species. *Annu. Rev. Anim. Biosci.* 6, 69–82. doi: 10.1146/annurev-animal-030117-014603

Burger, P. A. (2016). The history of Old World camelids in the light of molecular genetics. *Trop. Anim. Health Prod.* 48, 905–913. doi: 10.1007/s11250-016-1032-7

Cho, I.-C., Zhong, T., Seo, B.-Y., Jung, E.-J., Yoo, C.-K., Kim, J.-H., et al. (2011). Whole-genome association study for the roan coat color in an intercrossed pig population between Landrace and Korean native pig. *Genes Genomics* 33, 17–23. doi: 10.1007/s13258-010-0108-4

Daverio, M. S., Vidal-Rioja, L., Frank, E. N., and Di Rocco, F. (2017). Molecular characterization of the llama FGF5 gene and identification of putative loss of function mutations. *Anim. Genet.* 48, 716–719. doi: 10.1111/age.12616

David, V. A., Menotti-Raymond, M., Wallace, A. C., Roelke, M., Kehler, J., Leighty, R., et al. (2014). Endogenous retrovirus insertion in the KIT oncogene determines white and white spotting in domestic cats. *G3* 4, 1881–1891. doi: 10.1534/g3.114.013425

Dierks, C., Momke, S., Philipp, U., and Distl, O. (2013). Allelic heterogeneity of FGF5 mutations causes the long-hair phenotype in dogs. *Anim. Genet.* 44, 425–431. doi: 10.1111/age.12010

Dioli, M. (2016). Towards a rational camel breed judging: a proposed standard of a camel (*Camelus dromedarius*) milk breed. *J. Camel Pract. Res.* 23, 1–12. doi: 10.5958/2277-8934.2016.00001.1

Drögemüller, C., Rüfenacht, S., Wichert, B., and Leeb, T. (2007). Mutations within the FGF5 gene are associated with hair length in cats. *Anim. Genet.* 38, 218–221. doi: 10.1111/j.1365-2052.2007.01590.x

Druml, T., Dobretsberger, M., and Brem, G. (2015). The use of novel phenotyping methods for validation of equine conformation scoring results. *Animal* 9, 928–937. doi: 10.1017/S1751731114003309

Eshra, E. A., and Badawy, A. M. (2014). Peculiarities of the camel and sheep narial musculature in relation to the clinical value and the mechanism of narial closure. *Indian J. Vet. Anat.* 26, 10–13.

Farah, Z. (1993). Composition and characteristics of camel milk. *J. Dairy Res.* 60, 603–626. doi: 10.1017/S0022029900027953

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2016). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16, 314–324. doi: 10.1111/1755-0998.12443

Fontanesi, L., Tazzoli, M., Russo, V., and Beever, J. (2010). Genetic heterogeneity at the bovine KIT gene in cattle breeds carrying different putative alleles at the spotting locus. *Anim. Genet.* 41, 295–303. doi: 10.1111/j.1365-2052.2009.02007.x

Fontanesi, L., Vargiolu, M., Scotti, E., Latorre, R., Faussone Pellegrini, M. S., Mazzoni, M., et al. (2014). The KIT gene is associated with the english spotting coat color locus and congenital megacolon in checkered giant rabbits (*Oryctolagus cuniculus*). *PLoS One* 9:e93750. doi: 10.1371/journal.pone.0093750

Food Agriculture Organization of the United Nations (2013). *In Vivo Conservation of Animal Genetic Resources*. Rome: Food and Agriculture Organization of the United Nations. <doi>

Geissler, E. N., Ryan, M. A., and Housman, D. E. (1988). The dominant-white spotting (W) locus of the mouse encodes the *c-kit* proto-oncogene. *Cell* 55, 185–192. doi: 10.1016/0092-8674(88)90020-7

Groeneveld, L. F., Gregusson, S., Guldbrandtsen, B., Hiemstra, S. J., Hveem, K., Kantanen, J., et al. (2016). Domesticated animal biobanking: land of opportunity. *PLoS Biol.* 14:e1002523. doi: 10.1371/journal.pbio.1002523

Haase, B., Rieder, S., and Leeb, T. (2015). Two variants in the KIT gene as candidate causative mutations for a dominant white and a white spotting phenotype in the donkey. *Anim. Genet.* 46, 321–324. doi: 10.1111/age.12282

Hammond, A. (2007). *Saudi tribe holds camel beauty pageant. Oddly Enough*. Available at: https://www.reuters.com/article/us-saudi-camels-beauty-odd/saudi-tribe-holds-camel-beauty-pageant-idUSKUA74812720070427

Hauswirth, R., Jude, R., Haase, B., Bellone, R. R., Archer, S., Holl, H., et al. (2013). Novel variants in the KIT and PAX3 genes in horses with white-spotted coat colour phenotypes. *Anim. Genet.* 44, 763–765. doi: 10.1111/age.12057

Hébert, J. M., Rosenquist, T., Götz, J., and Martin, G. R. (1994). FGF5 as a regulator of the hair growth cycle: evidence from targeted and spontaneous mutations. *Cell* 78, 1017–1025. doi: 10.1016/0092-8674(94)90276-3

Higgins, C. A., Petukhova, L., Harel, S., Ho, Y. Y., Drill, E., Shapiro, L., et al. (2014). FGF5 is a crucial regulator of hair length in humans. *Proc. Natl. Acad. Sci. U.S.A.*111, 10648–10653. doi: 10.1073/pnas.1402862111

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108. doi: 10.1038/nrg1521

Holl, H., Isaza, R., Mohamoud, Y., Ahmed, A., Almathen, F., Youcef, C., et al. (2017). A frameshift mutation in KIT is associated with white spotting in the arabian camel. *Genes* 8:E102. doi: 10.3390/genes8030102

Jackling, F. C., Johnson, W. E., and Appleton, B. R. (2014). The genetic inheritance of the blue-eyed white phenotype in alpacas (*Vicugna pacos*). *J. Hered.* 105, 847–857. doi: 10.1093/jhered/ess093

Kadim, I. T., Mahgoub, O., and Purchas, R. W. (2008). A review of the growth, and of the carcass and meat quality characteristics of the one-humped camel (*Camelus dromedaries*). *Meat. Sci.* 80, 555–569. doi: 10.1016/j.meatsci.2008.02.010

Kagunyu, A. W., Matiri, F., and Ngari, E. (2013). Camel hides: production, marketing and utilization in pastoral regions of northern Kenya. *Pastoralism* 3:25. doi: 10.1186/2041-7136-3-25

Kahle, D., and Wickham, H. (2013). ggmap: spatial visualization with ggplot2. *R J.* 5, 144–161.

Karlsson, E. K., and Lindblad-Toh, K. (2008). Leader of the pack: gene mapping in dogs and other model organisms. *Nat. Rev. Genet.* 9:713. doi: 10.1038/nrg2382

Khalaf, S. (1999). Camel racing in the gulf. Notes on the evolution of a traditional cultural sport. *Anthropos* 94, 85–106.

Köhler-Rollefson, I. (1993). About camel breeds: a reevaluation of current classification systems. *J. Anim. Breed. Genet.* 110, 66–73. doi: 10.1111/j.1439-0388.1993.tb00717.x

Legrand, R., Tiret, L., and Abitbol, M. (2014). Two recessive mutations in FGF5 are associated with the long-hair phenotype in donkeys. *Genet. Sel. Evol.* 46:65. doi: 10.1186/s12711-014-0065-5

Lloyd-Jones, O. (1915). WHAT IS A BREED? Definition of word varies with each kind of livestock, and is based almost wholly on arbitrary decision of breeders—some strange contradictions—the meaning of "Pure-Bred"1. *J. Hered.* 6, 531–537. doi: 10.1093/oxfordjournals.jhered.a109038

Lyons, L. A., Imes, D. L., Rah, H. C., and Grahn, R. A. (2005). Tyrosinase mutations associated with Siamese and Burmese patterns in the domestic cat (*Felis catus*). *Anim. Genet.* 36, 119–126. doi: 10.1111/j.1365-2052.2005.01253.x

Mulsant, P., De Rochambeau, H., and Thébault, R. G. (2010). A note on linkage between the Angora and fgf5 genes in rabbits. *World Rabbit Sci.* 12, 1–6. doi: 10.4995/wrs.2004.585

Nazari-Ghadikolaei, A., Mehrabani-Yeganeh, H., Miarei-Aashtiani, S. R., Staiger, E. A., Rashidi, A., and Huson, H. J. (2018). Genome-wide association studies identify candidate genes for coat color and mohair traits in the iranian markhoz goat. *Front. Genet.* 9:105. doi: 10.3389/fgene.2018.00105

Omrani, A. S., Al-Tawfiq, J. A., and Memish, Z. A. (2015). Middle East respiratory syndrome coronavirus (MERS-CoV): animal to human interaction. *Pathog. Glob. Health* 109, 354–362. doi: 10.1080/20477724.2015.1122852

Ott, J., Wang, J., and Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* 16, 275–284. doi: 10.1038/nrg3908

Pallotti, S., Pediconi, D., Subramanian, D., Molina, M. G., Antonini, M., Morelli, M. B., et al. (2018). Evidence of post-transcriptional readthrough regulation in FGF5 gene of alpaca. *Gene* 647, 121–128. doi: 10.1016/j.gene.2018.01.006

Petersen, B.-S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D., and Franke, A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics* 18:14. doi: 10.1186/s12863-017-0479-5

Porter, V., Alderson, L., Hall, S. J. G., and Sponenberg, D. P. (2016). *Mason's World Encyclopedia of Livestock Breeds and Breeding*. Boston, MA: CABI, 1. doi: 10.1079/9781845934668.0000

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramadan, S., Nowier, A. M., Hori, Y., and Inoue-Murayama, M. (2018). The association between glutamine repeats in the androgen receptor gene and personality traits in dromedary camel (*Camelus dromedarius*). *PLoS One* 13:e0191119. doi: 10.1371/journal.pone.0191119

Schmidt-Nielsen, K. (1959). The physiology of the camel. *Sci. Am.* 201, 140–151. doi: 10.1038/scientificamerican1259-140

Teague, M. (2009). *Isn't She Lovely*. Washington, DC: National Geographic.

Volpato, G., Dioli, M., and Di Nardo, A. (2017). Piebald camels. *Pastoralism* 7:3. doi: 10.1186/s13570-017-0075-3

Wang, X., Cai, B., Zhou, J., Zhu, H., Niu, Y., Ma, B., et al. (2016). Disruption of FGF5 in cashmere goats using CRISPR/Cas9 results in more secondary hair follicles and longer fibers. *PLoS One* 11:e0164640. doi: 10.1371/journal.pone.0164640

Wardeh, M. F. (2004). Classification of the dromedary camels. *J. Camel Sci.* 1, 1–7.

Wong, A. K., Ruhe, A. L., Robertson, K. R., Loew, E. R., Williams, D. C., and Neff, M. W. (2012). A de novo mutation in KIT causes white spotting in a subpopulation of German Shepherd dogs. *Anim. Genet.* 44, 305–310. doi: 10.1111/age.12006

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Zelditch, M. L., Swiderski, D. L., Sheets, H. D., and Fink, W. L. (2004). *Geometric Morphometrics for Biologists*. San Diego, CA: Academic Press.

Zhang, L., He, S., Liu, M., Liu, G., Yuan, Z., Liu, C., et al. (2015). Molecular cloning, characterization, and expression of sheep FGF5 gene. *Gene* 555, 95–100. doi: 10.1016/j.gene.2014.10.036

Zhang, M. Q., Xu, X., and Luo, S. J. (2014). The genetics of brown coat color and white spotting in domestic yaks (Bos grunniens). *Anim. Genet.* 45, 652–659. doi: 10.1111/age.12191

Zhu, M., and Zhao, S. (2007). Candidate gene identification approach: progress and challenges. *Int. J. Biol. Sci.* 3, 420–427. doi: 10.7150/ijbs.3.420

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor and reviewer PO-tW declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

Copyright © 2019 Alhaddad and Alhajeri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 February 2019
doi: 10.3389/fgene.2019.00032

A Near Chromosome Assembly of the Dromedary Camel Genome

*Daniil Ruvinskiy[1], Denis M. Larkin[1,2] and Marta Farré[1,3]**

[1] Comparative Biomedical Sciences, Royal Veterinary College, University of London, London, United Kingdom, [2] The Federal Research Center, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, [3] School of Biosciences, University of Kent, Canterbury, United Kingdom

The dromedary camel is an economically and socially important species of livestock in many parts of the world, being used for transport and the production of milk and meat. Much like cattle and horses, the camel may be found in industrial farming conditions as well as used in sporting. Camel racing is a multi-million dollar industry, with some specimens being valued at upward of 9.5 million USD. Despite its apparent value to humans, the dromedary camel is a neglected species in genomics. While cattle and other domesticated species have had much attention in terms of genome assembly, the camel has only been assembled to scaffold level, which does not give a clear indication of the order or chromosomal location of sequenced fragments. In this study, the Reference Assistant Chromosome Assembly (RACA) algorithm was implemented to use read-pair information of camel scaffolds, aligned with the cattle and human genomes in order to organize and orient these scaffolds in a near-chromosome level assembly. This method generated 72 large size fragments (N50 54.36 Mb). These predicted chromosome fragments (PCFs) were then compared with comparative maps of camel and cytogenetic map of alpaca chromosomes, allowing us to further upgrade the assembly. This dromedary camel assembly will be an invaluable tool to verify future camel assemblies generated with chromatin conformation or/and long read technologies. This study provides the first near-chromosome assembly of the dromedary camel, thus adding this economically important species to a growing pool of knowledge regarding the genome structure of domesticated livestock.

Keywords: dromedary camel, genome, chromosome, assembly, camelids

OPEN ACCESS

Edited by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom

Reviewed by:
Igor V. Sharakhov, Virginia Tech, United States
Ernest Lam, Bionano Genomics, United States

****Correspondence:***
Marta Farré
mfarrebelmonte@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *29 October 2018*
Accepted: *17 January 2019*
Published: *05 February 2019*

Citation:
Ruvinskiy D, Larkin DM and Farré M (2019) A Near Chromosome Assembly of the Dromedary Camel Genome. Front. Genet. 10:32. doi: 10.3389/fgene.2019.00032

INTRODUCTION

Dromedary camels (*Camelus dromedarius*) are members of the Camelidae family, the only family with extant species of the suborder Tylopoda, part of the Cetartiodactyla order. Camelids appeared ~20 million years ago (Heintzman et al., 2015), and nowadays two main tribes of camelids exist, Old World camelids including the dromedary and Bactrian camel, and New World camelids with llamas, alpacas, vicunas, and guanacos. Camelids are characterized by karyotypes with a diploid number of $2n = 74$ and almost identical chromosomes, with only slight variations in the amount and distribution patterns of heterochromatin (Balmus et al., 2007). Dromedary camels, as well as other camelid species, are adapted to harsh environments with dry, arid conditions and high temperatures (Gebreyohanes and Assen, 2017). Dromedary camels no longer exist in the wild; however, they are widely farmed in many countries with hot climates, such as Egypt, Syria, Libya, Somalia, Ethiopia,

and Kazakhstan (Faye, 2015). Camels are not only used as means of transport, but also for dairy and meat production (Kebede et al., 2015). They are capable of producing milk for long periods of time and maintain its production under conditions where other animals would starve, thanks to having an unusually well-adapted udder for arid conditions (Alluwaimi et al., 2017). Although its economical and societal importance in developing countries, dromedary camel genomics has been understudied, and only recently, two dromedary camel genome assemblies were released (Wu et al., 2014; Fitak et al., 2016). However, both are assembled at scaffold level with an N50 of 4.1 and 1.40 Mb, respectively, making them unsuitable for in-depth use in evolutionary and applied genomics. To facilitate studies of genotype-to-phenotype associations for marker-assisted selection and breeding, high-quality chromosome-level assemblies are essential (Andersson and Georges, 2004). While such assemblies are established for popular livestock species, they are not available for those additional livestock species widely used in developing countries, including camels.

The African dromedary camel genome (Fitak et al., 2016) was assembled using next-generation sequencing (NGS) technologies. However, the use of short-read NGS data rarely produces assemblies at a similar level of integrity as those provided by traditional methodologies. NGS methods are incapable of generating long error-free contigs or scaffolds to cover chromosomes completely, requiring physical maps to upgrade NGS genomes to chromosome level (Lewin et al., 2009). Although new methodologies are being developed to overcome these limitations [e.g., long reads (Rhoads and Au, 2015), optical (Neely et al., 2011) or chromatin conformation maps (Lieberman-Aiden et al., 2009)], they often rely on hundreds of micrograms of high-molecular-weight DNA, which for some species are difficult to obtain, are usually expensive and suffer from misassembles. Bioinformatic approaches, e.g., the Reference-Assisted Chromosome Assembly (RACA) algorithm (Kim et al., 2013), were developed to approximate near chromosome-sized fragments for a *de novo* assembled NGS genome. RACA can assemble target genomes with no existing physical maps, utilizing their comparison to chromosome-level assemblies of reference and outgroup genomes, and read-pair data from target genome. RACA is suited for large, fragmented datasets such as the dromedary genome (Kim et al., 2013). Other reference-based algorithms e.g., RAGOUT (Kolmogorov et al., 2014) do not use the target assembly read-pair data to verify scaffold structures and orders, meaning that the target species-specific rearrangements could be missed from the reconstructed chromosome fragments, which could prove to be a problem in future candidate gene research, as a lower quality genome assembly will produce more false-negative and false-positive association signals, reducing the value of association studies (Goldfeder et al., 2016). Moreover, RACA has been successfully used for other genome assembly projects, including mammals [such as Tibetan antelope and red fox (Kim et al., 2013; Rando et al., 2018)] and birds [peregrine and saker falcons, ostrich, pigeon, and budgerigar (Damas et al., 2017; O'Connor et al., 2018)]. Finally, RACA assemblies could provide an independent source to prove and/or further improve assemblies produced with such methods as HiC, 10X or Dovetail Chicago.

In this report, therefore, we assembled the dromedary camel genome to near-chromosome level, using our previously established methodology (Damas et al., 2017). First, RACA was run to create predicted chromosome fragments (PCFs) and identify putatively chimeric scaffolds. These scaffolds could potentially contain structural errors and affect accuracy of PCFs or any other assemblies which would use them intact, therefore a subset of broken scaffolds was tested by polymerase chain reaction (PCR). Then, a second round of RACA was run to create a new, refined set of PCFs. And finally, taking advantage of the very stable camelid karyotypes (Balmus et al., 2007), we integrated previously published physical maps of dromedary camel (Balmus et al., 2007) and alpaca (Avila et al., 2014) to obtain a set of 72 chromosome fragments, with more than 80% of camel chromosomes assembled into three or less fragments. This new assembly will foster further genomic research into this special species and allow for improved genotype-to-phenotype studies.

MATERIALS AND METHODS

Using the Reference Assisted Chromosome Assembly (RACA) to Assemble the Dromedary Camel Genome

Reference Assisted Chromosome Assembly was used to further assemble the dromedary camel genome into PCFs (Kim et al., 2013). As inputs, RACA took a target species' (dromedary camel, Cdrom64K) scaffolds (Fitak et al., 2016), read-pair information, and the genome assemblies of a reference (cattle, bosTau6) and outgroup (human, hg19) species. The reference and outgroup species diverged 64.2 and 94.0 million years (MY) from camel, respectively.

Camel Read Sequence Data and Mapping

Sequence reads for dromedary camel (SRR2002493, SRR1950615, and SRR1693817) (Fitak et al., 2016) were downloaded from the National Center for Biotechnology Information (NCBI) using SRA toolkit v.2.8.2 (Leinonen et al., 2011). FastQC v.0.11.5 (Andrews, 2010) was used to evaluate the reads to decide on quality trimming. Bowtie2 v.2.3.0 (Langmead and Salzberg, 2012) was used to map camel reads to camel scaffolds, with insert minimum and maximum lengths of 250 and 750 bp for corresponding libraries (according to sequencing library information), trimming three base pairs from the 3′ end of each read.

Genome Alignments

To avoid spurious alignments, only original scaffolds longer than 10 Kb were used in this study. Lastz v.1.02.00 (Harris, 2007) was used for alignment of the camel scaffolds against the cattle assembly. Sequence alignments were concatenated into "chains," which were then transformed into hierarchical

"nets" alignments, according to alignment scores using Kent-library tools as described previously (Kent et al., 2003; Damas et al., 2017). The chain and net genome alignments between the human and cattle genomes were downloaded from the UCSC Genome Browser.

Reference Assistant Chromosome Assembly considers user-provided adjacencies of syntenic fragments (SFs) originating from different scaffolds as "reliable" and uses them to adjust read mapping thresholds. We defined reliable SF adjacencies *in silico*, using BLAT to map cattle genes to camel scaffolds. Cattle genes that mapped to two different SFs from two different camel scaffolds were then used as reliable SF adjacencies. These adjacencies were considered reliable, because if these SFs are not adjacent, the corresponding gene would need to be broken, which is unlikely due to high levels of gene conservation between mammalian genomes (Elsik et al., 2009).

RACA Run I

To improve the reliability of the final results, we ran RACA twice. Initially, the RACA algorithm was run to identify putatively chimeric scaffolds in the camel assembly, following our previous methodology (Farré et al., 2016). SFs were constructed at a 150 Kb resolution of SF detection, with default parameters except for: WINDOWSIZE = 10 and MIN_INTRACOV_PERC = 5.

PCR Testing of Putatively Chimeric Scaffolds

Primer pairs for testing putatively chimeric scaffolds were designed using Primer3 (v.2.3.6) (Untergasser et al., 2012) with optimum primer size of 20 bp (**Supplementary Table S3**). Only putatively chimeric scaffolds with a break interval size of <6 Kb were included in this analysis. Primers were chosen from camel sequences exhibiting high-quality alignments with the reference genome and the PCR product spanning the putatively chimeric join.

Camel DNA quality and concentration were tested using the Nanodrop 2000c (Thermo scientific). PCR was performed in a 10 μl volume with 5 μl Taq Polymerase Mix, 2 μl ddH_2O, 1 μl of each primer at 2 μM in ddH_2O and 1 μl of 30 ng/μl DNA solution. Thermal cycling was performed in the T100 thermal cycler (Bio-Rad) for 35 cycles: initial denaturation at 95°C for 3:00 min, 30 cycles of 95°C for 30 s (denaturation), 59–60°C at 1:00 min (annealing) and extension at 72°C at 1:00 min per PCR product 1,000 bp. Electrophoresis was done using the Sub Cell GT electrophoresis cell (Bio-Rad) with the power-pac basic power supply (Bio-Rad) with times ranging 20–40 min. PCR products were stained with SYBR-safe (Invitrogen) in a 1.5 and 1% agarose (Sigma) gel for PCR product lengths up to 2 and 4 Kb, respectively. Gels were visualized in a ChemiDOC MP system (Bio-Rad).

Polymerase chain reaction was done for two sets of primers per each putatively chimeric scaffold: the first set tested chimeric scaffold structure, and the second set tested the alternative (RACA-suggested) order of SFs from this scaffold, if a negative PCR result was observed for the first PCR following previous publication (Farré et al., 2016).

RACA Run II

Polymerase chain reaction confirmed non-chimeric scaffolds were included as an additional set of reliable SF adjacencies. The results of PCR testing also allowed to discern a physical coverage threshold of 212.5 read pairs (representing a coverage percentage of 51.16%), above which putatively chimeric scaffolds suggested by RACA are expected to be non-chimeric. As such, the second RACA was run with only one modified parameter: MIN_INTRACOV_PERC = 51.16. The results of the RACA run II were then transformed into a FASTA genome file, by joining the SFs in accordance with RACA's instructions.

Evaluating PCFs and Assigning Them to Chromosomes

Predicted chromosome fragments obtained in RACA run II were manually compared with the fluorescence *in situ* hybridization (FISH) comparative map of the dromedary camel and human genomes (Balmus et al., 2007). The alignment output of camel PCFs to human chromosomes generated by RACA was used to verify and order PCFs along camel chromosomes. In addition to this, and making use of the highly stable camelid karyotypes, we compared the PCFs to a published cytogenetic map of alpaca (*Lama pacos*) (Avila et al., 2014). Coding sequences (CDSs) of the gene markers used in the alpaca map were downloaded from NCBI and mapped to dromedary camel PCFs using BLAT with default parameters. Only alignments spanning more than 80% of the CDS were considered reliable and analyzed further. PCFs with at least one marker were assigned to dromedary camel chromosomes following the alpaca gene map, while PCFs with at least two markers in the same order as in Avila et al. (2014) were placed and oriented into camel chromosomes (Avila et al., 2014).

Finally, the Benchmarking Universal Single-Copy Orthologs tool (BUSCO) (Simão et al., 2015) with the mammalian and laurasiatherian databases was used to verify completeness of core genes in the assembly. We then used REAPR (Hunt et al., 2013) to identify errors in our genome assembly without the need for a reference sequence with the short-insert size libraries.

RESULTS

Following our previous publication (Farré et al., 2016), our approach to assemble the dromedary camel genome to near-chromosome level involved three steps: (1) the construction of PCFs using the RACA algorithm; (2) PCR and computational verification of a subset of scaffolds that might contain species-specific chromosome structures or be chimeric; and (3) creation of a refined set of PCFs using the verified scaffolds and adjusted parameters to run RACA. We then used previously published physical maps of dromedary camel (Balmus et al., 2007) and alpaca (Avila et al., 2014) to verify the PCFs and assign them to dromedary camel chromosomes.

Construction of PCFs From Scaffolds

A total of 4,922 camel scaffolds longer than 10 Kb, encompassing 1.99 Gb and representing 92.6% of the scaffold-based assembly, were aligned to cattle genome using lastZ and then concatenated

to chains and nets as previously described (Kent et al., 2003). Overall, pair-wise alignments spanned 98.75 and 99.50% of cattle chromosomes for camel-cattle and human-cattle pairs, respectively. Five dromedary camel pair-end read libraries were mapped to camel scaffolds using Bowtie2 and the mapping coverage for each library was calculated using bedtools (Quinlan and Hall, 2010). Only three libraries (SRR2002493, SRR1950615, and SRR1693817) had an average coverage >17x of the camel genome and were used to run RACA.

An important input file to train RACA consists of SF adjacencies with a prior knowledge of being connected. To create this file, we made use of the high gene structure conservation in mammalian species (Elsik et al., 2009) and assumed that genes in one species are highly likely to maintain their structure in another closely related species. Therefore, we mapped cattle genes to camel scaffolds to identify genes aligned to two SFs containing two different camel scaffolds. A total of 23,819 cattle genes were used, of which 50 mapped to two different camel scaffolds, and were included as reliable adjacencies. Overall, the initial RACA run resulted in 73 PCFs with an N50 of 54.36 Mb covering 94.0% of scaffold-based assembly (**Table 1**).

Reference Assistant Chromosome Assembly introduced 49 breaks in 46 (2.6%) camel scaffolds, and they were considered as putatively chimeric joints. These scaffolds contained structural differences from the cattle and human genomes, meaning that they could negatively affect PCF structures if proven to be chimeric. In order to assess these joints, primers were designed for 27 out of 49 putatively chimeric joints. A total of 14 of the 27 selected intervals resulted in PCR products of expected sizes, indicating that these joints were not chimeric (**Figure 1**, **Table 2**, and **Supplementary Table S1**). For joints with no amplification in PCR round I, we tested the alternative arrangements of SFs suggested by RACA (**Figure 2** and **Table 2**). If the order of SFs suggested by RACA was confirmed by PCR, the corresponding scaffold(s) were classified as chimeric.

TABLE 1 | Statistics for RACA-based assembly of dromedary camel genome.

Statistics	Scaffold assembly	RACA run I	RACA run II
No. scaffolds	4,922	1,797	1,797
No. PCFs	NA	73	72
Homologous to complete reference chromosomes	NA	5	6
Total length (% of original assembly)	1,998,420,525 (100%)	1,886,430,396 (94.4%)	1,886,430,696 (94.4%)
N50 (Mb)	1.40	54.36	54.36
Max. length (bp)	9,719,801	122,837,232	122,837,232
Min. length (bp)	10,001	206,422	206,422
*Max. no. scaffolds	NA	97	100
*Min. no. scaffolds	NA	1	1
No. broken scaffolds	NA	46 (2.60%)	47 (2.62%)

Min/max number of scaffolds are the minimum and maximum number of scaffolds represented in single PCFs.

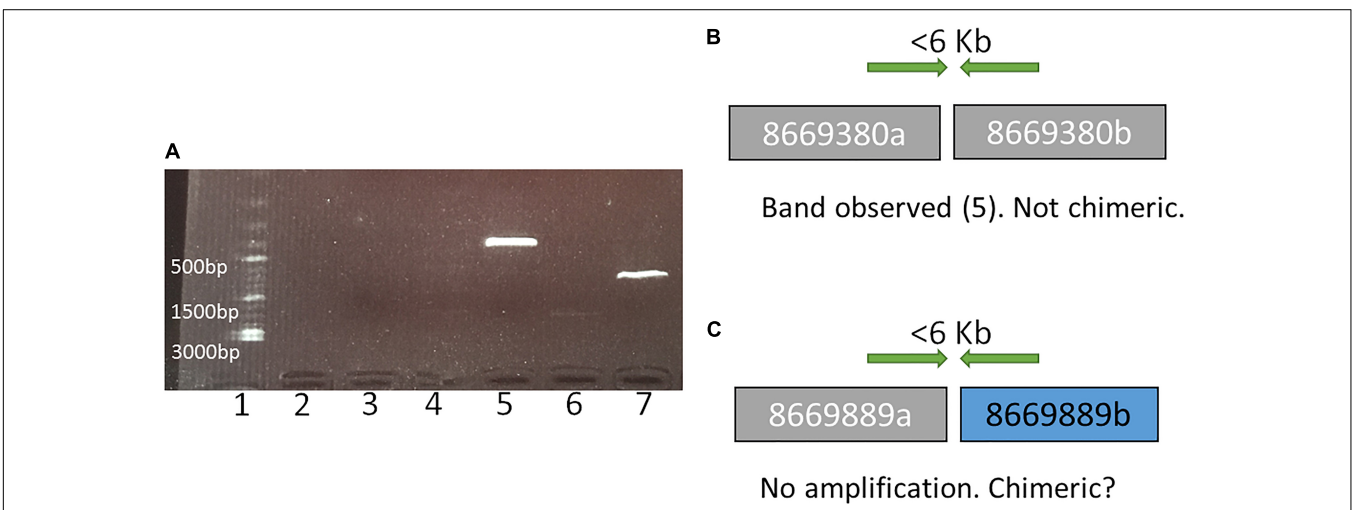


FIGURE 1 | An example of representation and interpretation of the results of polymerase chain reaction (PCR) run I. **(A)** an electrophoresis gel, where in well 1 is a DNA ladder, 2 is the control (no DNA), 3 and 4 exhibit no PCR amplification for tested SF adjacencies in scaffolds 8669889 and 8669009, respectively, broken by RACA; 5 and 7 show PCR amplification for tested SF adjacencies in scaffolds 8669380 and 8669417, respectively, broken by RACA; 6 shows PCR product of unexpected size (2000 bp) for scaffold 8667696. **(B)** A schematic representation and interpretation of the PCR results from well 5 (scaffold 8669380, SFs 8669380a and 8669380b); **(C)** A schematic representation and interpretation of the PCR results from well 3 (scaffold 8669889, SFs 8669889a, and 8669889b).

TABLE 2 | Verification of putatively chimeric scaffolds by PCR.

Statistics	Camel
Pair-end read physical coverage within scaffolds	5.5 – 329.7
No. split SF adjacencies by RACA (default param.)	49
No. tested scaffold split regions	27
No. amplified split regions (confirmed SF joints)	14
No. non-amplified split regions	13
No. tested RACA-suggested adjacencies	18
No. amplified adjacencies (chimeric SF joints)	7
No. non-amplified adjacencies	11
Final no. ambiguous SF joints from tested split regions	11
Selected pair-end read spanning threshold	212.5
No. tested split regions found below selected threshold	22
No. chimeric SF joints	7
No. confirmed SF joints	4
No. ambiguous SF joints	11
No. tested split regions found above selected threshold	10
No. chimeric SF joints	0
No. confirmed SF joints	10
No. ambiguous SF joints	0

TABLE 3 | Number of PCFs per camel chromosome.

No. PCFs	No. chromosomes	% chromosomes
1	12	33.3
2	13	36.1
3	4	11.1
>3	5	13.9
Unknown	2	5.6

Further details can be found in ***Supplementary Table S3****.*

This resulted in seven of 18 tested intervals being classified as chimeric (**Table 2** and **Supplementary Table S1**). The reason there were more tested structures in the second run of PCR than there were negative results in the first run, is because there were two alternative SF arrangements that could be tested in the second PCR round (one per flanking SF) and for some scaffolds we tested both arrangements. Overall, seven scaffolds were confirmed as chimeric, while 14 were shown to be real. We could not make any conclusions regarding six scaffolds corresponding to 11 SF adjacencies (**Table 2**), because no PCR products were amplified in either of the rounds.

To estimate which of the remaining split scaffolds (>6 Kb or with ambiguous PCR results) were likely to be chimeric, we empirically identified a genome-wide minimum physical coverage (Meyerson et al., 2010) level in the SFs joining regions for which (and higher) the PCR results were consistent with RACA predictions. A physical coverage threshold of 212.5x was established, which would allow us to identify additional putatively chimeric scaffolds without any additional scaffold verification (**Table 2** and **Supplementary Table S1**).

Construction of a Refined Set of PCFs

Polymerase chain reaction-verified scaffolds, confirmed as non-chimeric but with a physical coverage below the new set threshold were used as additional reliable adjacencies for RACA run II. This run resulted in a final set of 72 PCFs with an N50 of 54.36 Mb (**Table 1**). The total length of the RACA assembly was ∼1.89 Gb. The longest PCF spanned 122.84 Mb and included 74 scaffolds, while the shortest was 206 Kb in size, containing only one scaffold. Six PCFs were homologous to complete cattle chromosomes (BTA9, BTA12, BTA19, BTA24, BTA25, and

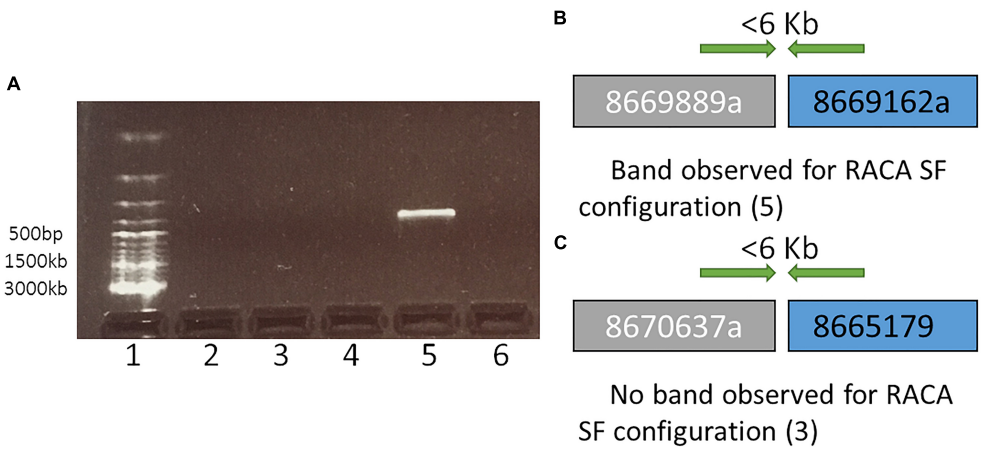


FIGURE 2 | A representation and interpretation of the results of PCR run II. **(A)** An electrophoresis gel, where in well 1 is a DNA ladder; 2 is the control (no DNA); 3, 4, and 6 show no amplification for tested SF adjacencies 8670637a and 8665179, 8667696a and 8667368, 8669739a and 8669686, respectively. Order of these SF was suggested by RACA; 5 exhibits amplification for adjacent SFs 8669889a and 8669162b in the order suggested by RACA. **(B)** A schematic representation and interpretation of the results form well 5 (SFs 8669889a and 8669162b were confirmed to be adjacent suggesting that the original scaffolds 8669889 and 8669162 these two SFs originate from were chimeric); **(C)** A schematic representation and interpretation of the results from well 3. The order of SFs 8670637a and 8665179, adjacent in RACA's output was not confirmed due to no PCR amplification, suggesting that no conclusion about the chimeric nature of scaffolds 8670637 and 8665179 could be made.

BTA27; **Figure 3**), from which only one (BTA19) showed an intrachromosomal rearrangement between cattle and dromedary genomes. A total of 46 scaffolds, representing 2.6% of scaffolds used by RACA, were still split despite some being present in reliable adjacencies.

Assessment of PCFs With Dromedary and Alpaca Cytogenetic Maps and Generation of a Final Chromosome Level Assembly

In order to verify the RACA assembly, we compared our PCFs to previously published physical maps for dromedary camel and alpaca. First, PCFs mapping to two or more human chromosomes were compared to the dromedary camel-human cytogenetic map (Balmus et al., 2007). A total of 61 PCFs, representing 87.8% of the total assembled genome, agreed with FISH, while six PCFs (12.2% of assembled genome) presented disagreements. Four of the PCFs that disagree with FISH data (PCFs 2a, 8b, 7a_10_20a and 10e_21a) contained a small fragment (<3 Mb of size) mapping to a human chromosome not revealed by FISH. However, these PCFs might be correct, since the sizes of the small fragments are below FISH resolution. Instead, PCF 17a mapped to two human chromosomes and the SFs were above FISH resolution, as such it was manually broken following human alignments in the regions with the lowest adjacency score produced by RACA. Finally, PCF24 was homologous to the entire human chromosome 18 (HSA18), but FISH data indicates that HSA18 corresponds to camel chromosomes 30 and 24. However, we were not able to separate the two fragments.

Then, taking into account the high karyotype stability in all camelid species (Balmus et al., 2007) we used the alpaca physical map (Avila et al., 2014) to assess the internal structure of the PCFs. A total of 52 alpaca genes successfully mapped to 26 camel PCFs (**Supplementary Table S2**). Although 12 PCFs contained only one gene of the set, it allowed us to confirm their correct placement into camel chromosomes. At least two genes mapped to 15 PCFs, allowing us to orient and assess their structure. Two PCFs (PCF 6b and 2c_3a_16a) disagreed with the alpaca gene map and were manually broken (**Supplementary Tables S2**, **S3**). By using the alpaca gene map with enough marker information, we identified these two more disagreements not detected with the FISH data only; therefore, by integrating two physical maps we produced a more reliable assembly (**Supplementary Figure S1**).

After verifying the PCFs and correcting the misassemblies, we used both physical maps to place and orient the PCFs into camel chromosomes (**Supplementary Table S3**). In doing so, more than 80% of chromosomes were assembled into three camel PCFs: 12 chromosomes were presented by a single PCF, 13 by two PCFs, and four by three PCFs (**Table 3**). Five camel chromosomes were represented by more than three PCFs, while two chromosomes (CDR24 and CDR30) remained within the PCF24 as we were not able to break it. Then, we assessed the assembly contiguity using the BUSCO (Simão et al., 2015) with two sets of orthologous genes (**Figure 4**). The newly improved assembly contains more complete single copy BUSCOs and less fragmented genes in both the mammalian

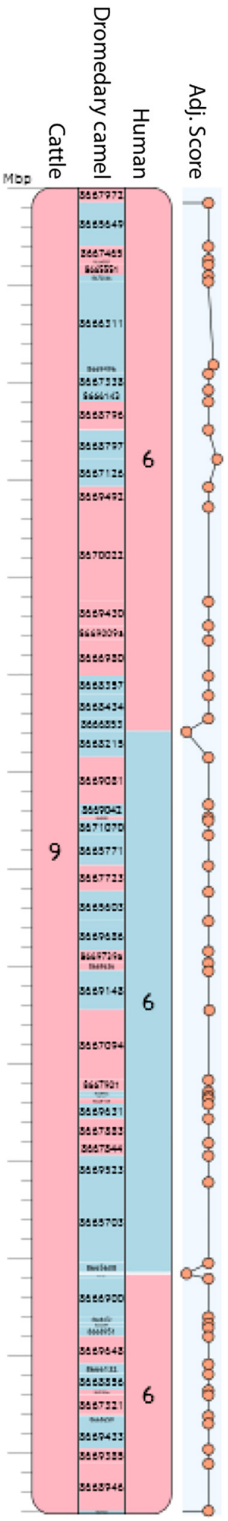


FIGURE 3 | Camel chromosome 8 corresponding to PCF 9. Blue blocks indicate positive (+) orientation of tracks compared with the camel chromosome while red blocks, negative (–) orientation. Numbers inside each block represent cattle and human chromosomes or dromedary scaffold IDs. Adjacency scores are shown on the right-hand side of the PCF. The rest of the chromosomes can be found in **Supplementary Figure S1**.

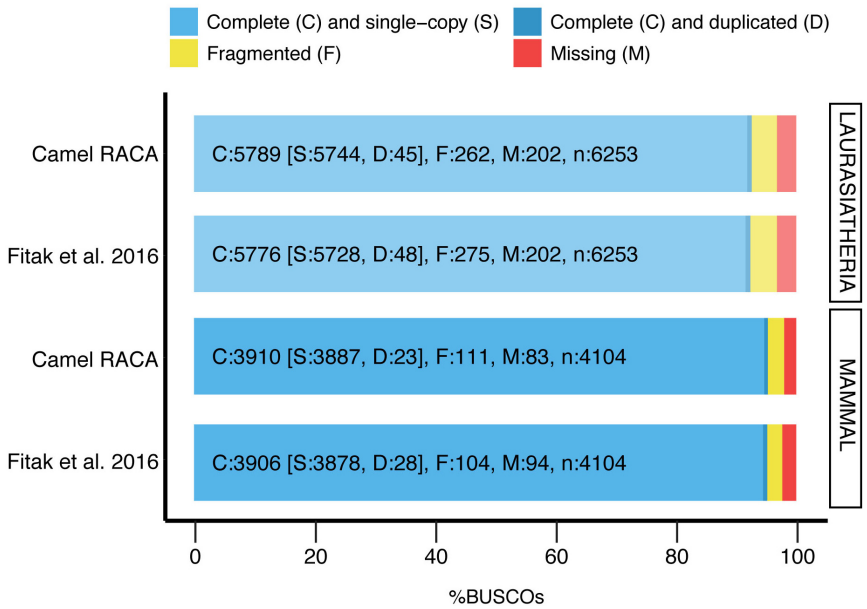


FIGURE 4 | Genome assembly evaluation. The BUSCO dataset of the mammalia odb9 including 4,104 BUSCOs, and the laurariatheria odb9, with 6,253 BUSCOs, were used to assess the new assembly (camel RACA) and the original scaffold-based assembly (Fitak et al., 2016).

and laurariatherian sets, showing an increase of contiguity. Finally, REAPR (Hunt et al., 2013) was used to identify assembly errors without the need of a reference genome (**Supplementary Table S4** and **Supplementary Figure S2**). Overall, we achieved a final dromedary camel chromosome-level assembly by combining *in silico* reconstructions with physical maps.

DISCUSSION

In this study, we upgraded the previously published fragmented dromedary camel genome assembly to nearly chromosome-level using a combination of *in silico* chromosome reconstructions, PCR-verification and supporting data from camel and alpaca physical maps. This approach has been previously applied to mammalian genomes, such as the Tibetan antelope (Kim et al., 2013) ($2n$ = 60), red fox (Rando et al., 2018) ($2n$ = 34), and avian species, including pigeon and peregrine falcon (Damas et al., 2017), and showed high consistency when compared with third-generation sequencing methodologies (Holt et al., 2018). Our approach resulted in a remarkable reduction in fragmentation of the original dromedary assembly by 25-fold, and an N50 increase 35-fold. Compared to other mammalian genomes assembled using the same approach, RACA produced 72 PCFs for dromedary camel, while 60 and 128 PCFs were obtained for Tibetan antelope and red fox, respectively (Kim et al., 2013; Rando et al., 2018). These differences could be explained by three main factors, the initial fragmentation of the scaffold-based assembly, the choice of reference genome and the chromosome rearrangement rate of the phylogenetic clade. Dromedary camel original assembly has an N50 of 1.40 Mb, while Tibetan antelope scaffold N50 was 2.76 Mb, indicating that a higher N50 of the input assembly could reduce the number of PCFs obtained by RACA. Moreover, the divergence time between the Tibetan antelope and the chosen reference genome (cattle) is 24 MY, whereas the divergence time between dromedary camel and cattle is 64.2 MY, suggesting that choosing a reference closely related to the target species improves continuity of RACA assemblies. But this hypothesis does not hold for red fox results, since the fox scaffold-based assembly had an N50 of 11.8 Mb and dog was used as reference genome (with 14 MY divergence time). However, canid lineage is characterized by a high chromosome rearrangement rate including multiple chromosome fissions (Graphodatsky et al., 2000); while cetartiodactyl clade, specially camelids, show a more stable karyotype (Balmus et al., 2007). For RACA, greater similarity between genome structures of the target and reference genomes clearly improves PCF assembly. Thus, a way to further improve the camel assembly would be to use a phylogenetically closer reference genome, e.g., the alpaca genome currently being assembled.

Although the RACA and PCR approach produces reliable assemblies when compared to third generation sequencing methodologies (Holt et al., 2018), we validated the PCFs using previously published physical maps of FISH using human probes on camel chromosomes (Balmus et al., 2007) and alpaca gene mapping (Avila et al., 2014). Our PCF assembly, FISH map, and alpaca marker genes map were highly consistent, with only

eight discrepancies, four of which were too small to be detected by FISH (<3 Mb) and did not contain any marker genes. Only two disagreements were above FISH resolution and guided by FISH and alpaca marker genes we corrected one of them. The remaining one consisted of a PCF orthologous to the entire HSA18 and BTA24. However, as shown by FISH and the alpaca gene map, HSA18 is orthologous to two camel chromosomes (CDR24 and CDR30), but we were not able to split it because not enough marker genes from the alpaca set mapped to this PCF. Therefore, comparing PCFs to such data was important, because it allowed us to check whether our assembly was consistent with independent FISH results, perform further verification, and order PCFs along camel chromosomes.

Although placing the PCFs into chromosomes is important to the usability of the dromedary camel genome, more work is required to improve it further. Integrating spatial and sequence information simultaneously, by using Hi-C (Lieberman-Aiden et al., 2009) and/or optical mapping will resolve the inconsistencies we found between FISH and PCFs as well as assemble the PCFs into complete chromosomes. Moreover, sequencing technologies being able to resolve repetitive regions [such as PacBio and Oxford Nanopore (Jain et al., 2018)] will greatly improve the assembly and close the remaining gaps. However, all these approaches are expensive and might not be within the reach of communities working with livestock species in developing countries. Furthermore, the new approaches are not free from limitations, e.g., HiC could result in false rearrangements to be introduced within chromosomes or even errors in joining chromosomes together. Our assembly, therefore, could be used to flag such inconsistently assembled regions and eventually help resolving them. That is why our improved dromedary camel genome assembled at nearly chromosome level is a step forward to a high-quality camel assembly. Moreover, it will facilitate efficient association of phenotype to genotype studies (Glazer et al., 2015) fostering genomic research in camelid species and also inform research on evolution and speciation through chromosomal changes. Furthermore, the methodology used in this study is significantly cheaper compared to many NGS sequencing methods, allowing for lower-income projects to participate in research.

AUTHOR CONTRIBUTIONS

DR performed the analysis and drafted the manuscript. DR and MF interpreted the results. DR, DL, and MF wrote the final version of the manuscript.

FUNDING

This work was funded by the Biotechnology and Biological Sciences Research Council grant BB/P020062/1 (DL) and Russian Foundation for Basic Research (RFBR) grant 17-00-00147 (DL).

ACKNOWLEDGMENTS

We would like to thank Dr. Pamela Burger from the Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna for providing us with camel DNA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00032/full#supplementary-material

FIGURE S1 | Camel chromosomes. Blue blocks indicate positive (+) orientation of tracks compared with the camel chromosome while red blocks, negative (–) orientation. Numbers inside each block represent cattle and human chromosomes or dromedary scaffold IDs. Adjacency scores are shown on the right-hand side of the PCF.

FIGURE S2 | Dotplot showing the alignment of our new assembly compared to a previous dromedary camel assembly (Wu et al., 2014).

TABLE S1 | Polymerase chain reaction results and decision made regarding putative chimeric joints in dromedary camel assembly.

TABLE S2 | Alpaca genes mapped in dromedary camel PCFs.

TABLE S3 | Placement of PCFs into dromedary camel chromosomes on the basis of FISH and BAC markers.

TABLE S4 | Assessment of the quality of the new assembly using REAPR.

REFERENCES

Alluwaimi, A. M., Al Mohammad Salem, K. A., Al Ashqer, R. A., and Al shubaith, I. H. (2017). The camel's (*Camelus Dromedarius*) mammary gland immune system in health and disease. *Adv. Dairy Res.* 5, 1–6. doi: 10.4172/2329-888X.1000171

Andersson, L., and Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nat. Rev. Genet.* 5, 202–212. doi: 10.1038/nrg1294

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/

Avila, F., Baily, M. P., Perelman, P., Das, P. J., Pontius, J., Chowdhary, R., et al. (2014). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*Lama pacos*). *Cytogenet. Genome Res.* 144, 196–207. doi: 10.1159/000370329

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C., Alkalaeva, E. S., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative Cetartiodactyla ancestral karyotype. *Chromosom. Res.* 15, 499–515. doi: 10.1007/s10577-007-1154-x

Damas, J., O'Connor, R., Farré, M., Lenis, V. P. E., Martell, H. J., Mandawala, A., et al. (2017). Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res.* 27, 875–884. doi: 10.1101/gr.213660.116

Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324, 522–528. doi: 10.1126/science.1169588

Farré, M., Narayan, J., Slavov, G. T., Damas, J., Auvil, L., Li, C., et al. (2016). Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biol. Evol.* 8, 2442–2451. doi: 10.1093/gbe/evw166

Faye, B. (2015). Role, distribution and perspective of camel breeding in the third millennium economies. *Emir. J. Food Agric.* 27, 318–327. doi: 10.9755/ejfa.v27i4.19906

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2016). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16, 314–324. doi: 10.1111/1755-0998.12443

Gebreyohanes, M. G., and Assen, A. M. (2017). Adaptation mechanisms of camels (*Camelus dromedarius*) for desert environment: a review. *J. Vet. Sci. Technol.* 8, 1–5. doi: 10.4172/2157-7579.1000486

Glazer, A. M., Killingbeck, E. E., Mitros, T., Rokhsar, D. S., and Miller, C. T. (2015). Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3* 5, 1463–1472. doi: 10.1534/g3.115.017905

Goldfeder, R. L., Priest, J. R., Zook, J. M., Grove, M. E., Waggott, D., Wheeler, M. T., et al. (2016). Medical implications of technical accuracy in genome sequencing. *Genome Med.* 8:24. doi: 10.1186/s13073-016-0269-0

Graphodatsky, A. S., Yang, F., O'Brien, P. C. M., Serdukova, N., Milne, B. S., Trifonov, V., et al. (2000). A comparative chromosome map of the Arctic fox, red fox and dog defined by chromosome painting and high resolution G-banding. *Chromosom. Res.* 8, 253–263. doi: 10.1023/A:1009217400140

Harris, R. S. (2007). *Improved Pairwise Alignment of Genomic DNA*. Ph.D. thesis, Pennsylvania State University, State College, PA.

Heintzman, P. D., Zazula, G. D., Cahill, J. A., Reyes, A. V., MacPhee, R. D., and Shapiro, B. (2015). Genomic data from extinct North American camelops revise camel evolutionary history. *Mol. Biol. Evol.* 32, 2433–2440. doi: 10.1093/molbev/msv128

Holt, C., Campbell, M., Keays, D. A., Edelman, N., Kapusta, A., Maclary, E., et al. (2018). Improved genome assembly and annotation for the rock pigeon (*Columba livia*). *G3* 8, 1391–1398. doi: 10.1534/g3.117.300443

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* 14:R47. doi: 10.1186/gb-2013-14-5-r47

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi: 10.1038/nbt.4060

Kebede, S., Animut, G., and Zemedu, L. (2015). *The Contribution of Camel Milk to Pastoralist Livelihoods in Ethiopia: An Economic Assessment in Somali Regional State*. London: IIED. Available at: http://pubs.iied.org/pdfs/10122IIED.pdf

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11484–11489. doi: 10.1073/pnas.1932072100

Kim, J., Larkin, D. M., Cai, Q., Asan, Zhang, Y., Ge, R.-L., et al. (2013). Reference-assisted chromosome assembly. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1785–1790. doi: 10.1073/pnas.1220349110

Kolmogorov, M., Raney, B., Paten, B., and Pham, S. (2014). Ragout – a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* 30, I302–I309. doi: 10.1093/bioinformatics/btu280

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Lewin, H. A., Larkin, D. M., Pontius, J., and O'Brien, S. J. (2009). Every genome sequence needs a good map. *Genome Res.* 19, 1925–1928. doi: 10.1101/gr.094557.109

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11, 685–696. doi: 10.1038/nrg2841

Neely, R. K., Deen, J., and Hofkens, J. (2011). Optical mapping of DNA: single-molecule-based methods for mapping genomes. *Biopolymers* 95, 298–311. doi: 10.1002/bip.21579

O'Connor, R. E., Farré, M., Joseph, S., Damas, J., Kiazim, L., Jennings, R., et al. (2018). Chromosome-level assembly reveals extensive rearrangement in saker falcon and budgerigar, but not ostrich, genomes. *Genome Biol.* 19:171. doi: 10.1186/s13059-018-1550-x

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rando, H. M., Farré, M., Robson, M. P., Won, N. B., Johnson, J. L., Buch, R., et al. (2018). Construction of red fox chromosomal fragments from the short-read genome assembly. *Genes* 9:E308. doi: 10.3390/genes9060308

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3222. doi: 10.1093/bioinformatics/btv351

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40:e115. doi: 10.1093/nar/gks596

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Ruvinskiy, Larkin and Farré. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

OPINION
published: 12 March 2019
doi: 10.3389/fgene.2019.00167

Genetic Improvement in Dromedary Camels: Challenges and Opportunities

Mohammed A. Al Abri[1]* *and Bernard Faye*[2]

[1] Department of Animal and Veterinary Sciences, Sultan Qaboos University, Muscat, Oman, [2] UMR SELMET, CIRAD-ES, Montpellier, France

Keywords: dromedary camels, genetic improvement, challenges, opportunities, food security, climate change

INTRODUCTION

Adaptation to a hotter climate is vital for future livestock as heat stress can extremely reduce their productivity, health, and fertility (Hayes et al., 2013). Camels have developed, through millennia, the ability to produce quality meat, milk, and fiber in some of the hottest and most hostile environments in the globe. According to the FAO live animals statistics, the worldwide camel population is ∼35 million heads (FAO, 2019), most of which are in Somalia, Sudan, Niger, Kenya, Chad, Ethiopia, Mali, Mauritania, and Pakistan. Moreover, partly due to climatic changes, areas of camel rearing are expanding, especially in Africa (Faye et al., 2012). Among the large camelids (dromedary and Bactrian), dromedary camels compose about 95% of the population (Bornstein and Younan, 2013). Due to their unique physiology and in light of the current climate change impacts on ecosystems, camels are poised to be an excellent candidate species for production (Hoffmann, 2010). This is specifically true in regions where agro-pastoralism is being replaced by pastoralism due to climate change (Bornstein and Younan, 2013). However, to harness their potential, an improved understanding of the genetics underlying their unique biology is needed.

OPPORTUNITIES

The term "Livestock Revolution" was coined to describe the projected increase in demand for animal products due to population growth, increased income, and urbanization in developing countries. For example, demand for beef and milk is expected to rise to 2.7–30 million metric tons, respectively by the year 2020 (Delgado et al., 1999). Most camels are in developing countries and can contribute in meeting meat and milk demands if utilized efficiently. Currently, most of that demand in many Middle East and North Africa (MENA) countries is met either by importation or local production using commercial exotic livestock not adapted neither to local climatic conditions and low input systems dominating the region. Camels can not only contribute in boosting food security but also in job creation, poverty alleviation and economic diversification. Utilization of camels in production will also reduce their destructive impact on the environment as is the situation in Australia (Saalfeld and Edwards, 2010). There, camels have contributed to the reduction of vegetation not only due to the increase in sheer numbers but also because they can browse and graze on a wide range of plants that are avoided by or are inaccessible to other livestock such as thorny bushes (Stiles, 1988; Faye, 2011; Al-Jassim and Sejian, 2015).

Beside their adaptation to harsh environments, camels are multipurpose animals used for milk and meat production, hair/felt, racing, transportation, and tourism. Camels also have a slow metabolism which results in comparatively less feed requirements compared to other ruminant livestock. As a result, they produce less methane on the basis of body mass index (Dittmann et al., 2014). Moreover, camels' milk and meat are highly nutritional and are comparable and sometimes

OPEN ACCESS

Edited by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom

Reviewed by:
Hans D. Daetwyler, La Trobe University, Australia
Joram Mwashigadi Mwacharo, International Center for Agriculture Research in the Dry Areas (ICARDA), Ethiopia

***Correspondence:**
Mohammed A. Al Abri
abri1st@squ.edu.om

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *03 October 2018*
Accepted: *15 February 2019*
Published: *12 March 2019*

Citation:
Abri MAA and Faye B (2019) Genetic Improvement in Dromedary Camels: Challenges and Opportunities. Front. Genet. 10:167. doi: 10.3389/fgene.2019.00167

deemed better than cattle beef and milk. For instance, camel meat contains less fat than lamb or beef (Kadim et al., 2008) and its protein quality, assessed by the index of essential amino-acids in meat, is the highest among red meat (Raiymbek et al., 2015). Its milk contains between 3 and 10 times more vitamin C than cows' milk (Faye et al., 1997; Konuspayeva et al., 2009). It also contains lower β-casein and no β-lactoglobulin resulting in its hypo-allergic property (Konuspayeva et al., 2009). During the last decade, demand for camel milk and meat products have increased both locally (in arid regions) and internationally with products varying from milk and its derivatives to beauty products to hump fat. Thus, a number of camel intensive dairy farms have been established worldwide and are currently supplying local and international markets (Gossner et al., 2014). There is, therefore, a slow but steady integration of camels' products into national and global economies (Faye, 2018a).

However, to utilize camels' potential, they need to undergo genetic improvement while sustaining their genetic diversity. Examples of successful genetic improvement of production traits in other livestock species are plenty and have considerably reshaped the livestock industry worldwide. For example, pigs are now 25% leaner and grow faster today than 20 years ago (Rothschild and Plastow, 2008) and milk production of Holstein Friesian cattle saw an increase of 40–80 kg/cow/year between 1980 and 2010 (Hayes et al., 2013). Similar success stories are evident in poultry and beef cattle and, together, have resulted in cheaper and more abundant animal derived proteins being available to consumers. Through multi-trait genetic improvement programs, not only production traits can be improved, but also health traits such as resistance to Peste des petits ruminants (PPR) virus or Rift Valley fever (RVF) both of which can have devastating effects on camel health. In addition, genetic improvement can also target other commercially important traits such as racing ability, beauty (Faye, 2015) or ease/suitability for machine milking (Ayadi et al., 2013).

Relatively few studies have investigated the genetic variability of production traits in camels (Dioli, 2016; Hemati et al., 2017). However, the few studies that have been carried out so far indicated that camels have a high genetic variability which is due to the lack of selection and the current and historical movements of camels between countries for trade and sometimes war (Almathen et al., 2016). This variability was reflected in the heritabilities of various traits, indicative of the potential for ample genetic gain if systematic selection is to be implemented. For instance, heritability estimates of body weight and growth rates were moderate to high, 0.24–0.40, respectively (Al-Sobayil et al., 2006). In another study, heritability estimates for birth weight was 0.37 and that of average daily gain ranged between 0.25 and 0.49 (Almutairi et al., 2010). Also, the heritability estimates for milk yield at 305 days and test day yields were 0.24 and 0.22, respectively (Almutairi et al., 2010). Together, these heritabilities show that the respective traits can indeed be improved through selection.

Genetic improvement in camels can be pursued using various methods. The first is single gene tests currently incorporated into selection programs of other livestock (Rothschild, 2004). However, to our knowledge, apart from color coat genes (Almathen et al., 2018), no other traits have been mapped in camels in which single test genes can be developed for. The second is traditional genetic selection using Best Linear Unbiased Prediction (BLUP) to estimate Estimated Breeding Values (EBVs) using phenotypic and pedigree information (Henderson, 1984). A variation of this method is using genomic relationships (using molecular markers) instead of pedigree information (Rodríguez-Ramilo et al., 2015). The third is using Genomic Selection (GS) which calculates Genomic EBVs termed "GEBVs" (Meuwissen et al., 2001). GEBVs are calculated as the sum of the effects of genetic markers across the entire genome of each animal (Hayes et al., 2009). This method requires that the genetic marker effects be inferred from individual single nucleotide polymorphism (SNPs) on a large reference population with phenotypic information. Once these effects have been calculated, only marker information is required to calculate GEBV in later generations (Hayes et al., 2009). GS or BLUP using genomic relationships are thus most likely to be adopted in camel genetic improvement especially because camels are not traditionally pedigreed. GS is specifically recommended for camels due to their long generation intervals and can accelerate the rate of genetic gain compared to conventional selection schemes. Unlike in small ruminants where the generation interval is short and a cost benefit analysis has to justify the implementation of GS (Mrode et al., 2018), in large ruminants such as camels and cattle, the high benefits of GS are clear with higher genetic gains and profits as a result of the reduction in generation intervals (Konig et al., 2009). Additionally, GS can result in increased accuracies of EBVs for young bulls and reduces the cost of progeny testing. In later generations, when more pedigree and phenotypic data become available, GS can be combined with individual and progeny phenotypic information in selection schemes. Moreover, accurate parentage testing can be obtained as a byproduct of genotyping animals for GS. A limitation in implementing GS however is the cost of genotyping, although that can be mitigated by using low density SNP panels (Abo-Ismail et al., 2018) or genotyping only a fraction of the genome, using Restriction-associated DNA (RAD) sequencing (Kess et al., 2016) or Genotyping by sequencing (GBS) (Elshire et al., 2011).

There is indeed an immediate potential in the existing camel dairies worldwide to ignite the spark of camel genetic improvement as they are consistent in pedigree and phenotypic collection. The different farms in Saudi Arabia (SA), United Arab Emirates (UAE), Kenya, and Bahrain can be the starting point for a genetic improvement in dairy camels if they participate in a common genetic evaluation program. To our knowledge, little communication and collaboration is currently practiced between these dairies, due primarily to competition. However, this lack of collaboration is bound to fade away with the realization that cooperation will improve long term profitability. Under such cooperation, records pertaining to milk production and health traits can be exchanged between the dairies as well as verified pedigree data. This exchange can help create a virtually common nuclear flock which can be utilized for traditional genetic evaluation of sires and dams. At a later stage, genomic selection can be practiced in order to speed up the genetic gain.

In addition, genetics of the elite animals can be disseminated to camel owners in respective countries. The realized genetic gain in the camel owners' herds shall encourage them to participate in genetic improvement programs. This will increase the number of participatory herds and the genetic variability accessible to the genetic evaluation program and accelerate genetic improvement. As the numbers of herds increase and more pedigree data become available, the evaluation can be extended to other traits such as beef, racing, and beauty. That in turn will help with the classification of the camel population into beef and dairy individuals and the identification of elite individuals in each category. This classification and will later make it easier for investors and owners to make future breeding decisions and reinforce the industry. An alternative to starting with the camel dairies for genetic improvement is starting with the camel owners themselves by forming cooperative community based breeding programs. These can begin with a nuclear flock formed by the owners that expand to include more owners in future. Such programs are found in developing countries for sheep and goats (Wurzinger et al., 2011) and have been successfully implemented in small ruminants (Gizaw et al., 2014; Mrode et al., 2018). This is, however, a more challenging approach and requires more upfront investment mostly by the funding agencies. In order to reduce the running cost, this approach needs to make use of modern digital systems such as mobile phones or tablets for recording performance and pedigree data (Mrode et al., 2016) and perhaps novel technologies such as automated monitoring systems which are now successfully used in dairy cattle (Stangaferro et al., 2016).

CHALLENGES

Despite its unique potential and increased contribution to food security, comparatively less attention has been paid to camels compared to other livestock species (Faye, 2015). Camels' genetics and genomics research is not an exception to this trend. Consequently, there are relatively few published studies in the area of camel genetics and genomics albeit ongoing research efforts (Jirimutu et al., 2012; Burger and Palmieri, 2014; Al-Swailem et al., 2018) notably through the International Camel Consortium for Genetic Improvement and Conservation (ICC-GIC) initiative. This is due, in part, to the lack of genomics tools to conduct such studies. For instance, the camel reference genome has not yet been released and no commercial genotyping platform has been developed for the species. Such platforms can be used to discover QTLs with impacts on specific traits using Genome Wide Association Studies (GWAS) and are the main engine for GS programs. Thus, whilst many QTLs have been reported using GWAS in sheep, cattle, and horses, none have been reported in camels. For example, endurance racing in Arabian horses was found to be partially controlled by 5 QTLs (Ricard et al., 2017) while in thoroughbred racing horses, a single mutation in the myostatin gene (*MSTN*) was found to profoundly affect the racing speed and stamina (Bower et al., 2012). Also, in cattle, variations in the *FABP4* gene were found to be significantly associated with milk yield and milk protein percentage (Zhou et al., 2015). Additionally, with the exception of dairy camels and to a less extent in racing in Dubai, very limited traditional genetic selection is applied (Faye, 2015).

Moreover, countries harboring most of the camel population are in different development stages pertaining to agriculture and infrastructure development. Thus, creation of intensive or peri-urban camel dairy or beef industries requires immense infrastructure investments, support and coordination between all stakeholders all of which are challenging. Although there is a gradual urbanization of some of the pastoral camel populations (Faye, 2015), most of the camel populations are still under traditional farming systems. As a result most camels do not possess unique identification number which hampers pedigree recording, good farm management, and performance recording (Caja et al., 2013). The relatively small herd size and scattered herds further complicate this issue making it difficult and costly to collect phenotypic data.

Another challenge facing the genetic improvement in camels is difficulty in disseminating superior genetics due to the difficulty of performing Artificial Insemination (AI). This is due primarily to the difficulty in semen collection and handling (due to the gelatinous nature of seminal plasma). In addition, deep freezing of camel semen has proved to be highly a challenge. Although research groups have tried different buffers and diluents as mediums for freezing camel semen (Skidmore et al., 2013), to date, it remains a challenge facing AI in camels. Moreover, unlike cattle, female camels are induced ovulators i.e., the females need to be induced to ovulate prior to AI (El-Bahrawy, 2018). While it is possible to use GnRH for inducing ovulation in camels, it depends on the stage of follicular development and/or estrous cycle (Manjunatha et al., 2015). A promising protocol for timed breeding called FWsynch in which a GnRH and PGF2α based hormonal regimen to synchronize the follicular wave was recently developed with satisfactory results (Manjunatha et al., 2015). However, while the cost of implementing such timed breeding regimens can be justifiable by research centers and camel dairies, they may not be as such for many camel owners specially that most of them reside in remote areas.

In the era of genomics, phenotypes are still very important and the availability of accurate and well defined phenotypes to be used in genetic studies and evaluation programs is imperative (Gonzalez-Recio et al., 2014). Unlike in developed countries, most of the camel herds in developing countries lack breed societies. They also do not have on farm automated milk recording systems and do not collect health or fertility traits (Faye, 2018b). Therefore, phenotypic recording is seldom practiced in camel populations except in intensive dairy farms, research, or racing. This creates an obstacle for genetic improvement programs and would require a serious collaboration of owners and stakeholders to circumvent. If camel breeds are sometimes described at a national level, as for example in Saudi Arabia (Abdallah and Faye, 2012), Tunisia (Chniter et al., 2018), or Algeria (Oulad-Belkhir et al., 2013), there is no standardization of the traits and parameters to be systematically recorded. For example, despite proposal on linear scoring for udder morphology, there is no application at a large-scale recording system (Ayadi et al., 2016).

The final hurdle is that, in developing countries, camels' meat and milk products are generally more expensive than imported milk and beef or those produced locally (by advanced genetic stock from developed countries). This is expected given the cost of production and the lack of genetic improvement in camels. It is therefore challenging for small scale producers to survive without government subsidies and support. To increase the market share and potential for such producers, added value products (such as flavored milk, dry milk, cheese, sour milk, camel burgers, and sausages) need to be produced and smart marketing strategies need to be adopted. Such strategies could include awareness campaigns of the health benefits of camel products, attractive product packaging, online marketing and partnership with existing cattle dairies and beef production firms for distribution and marketing. Camel milk can be marketed as a functional food, optimal for infants and elderly (Nikkhah, 2011). Focus can be made on the antimicrobial, antioxidants, and antidiabetic components of camel milk (Hailu et al., 2016). All of this can increase the value of camel products and hence improve producers' profitability and alleviate their dependence on the governments in the long term. If producers' profitability improved, it would become more feasible for them to participate in genetic selection programs. Selection for economically important traits can be practiced and would reduce the production cost, thereby reducing prices and increasing long term competitiveness.

In conclusion, camels have a large potential that is underutilized due to technical, logistic, political, and economic challenges. However, these challenges are not insurmountable, and much can be done to exploit the camels' potential. Genetic improvement is certainly promising in camels but would require the collaboration of all stakeholders and deeper understanding of the potential of this exceptional animal.

AUTHOR CONTRIBUTIONS

MA conceived the idea and wrote the manuscript. BF participated in writing and reviewing the manuscript.

REFERENCES

Abdallah, H. R., and Faye, B. (2012). Phenotypic classification of Saudi Arabian camel (*Camelus Dromedarius*)by their body measurements. *Emirates J. Food Agric.* 24, 272–280. Available online at: https://www.ejfa.me/index.php/journal/article/view/871

Abo-Ismail, M. K., Lansink, N., Akanno, E., Karisa, B. K., Crowley, J. J., Moore, S. S., et al. (2018). Development and validation of a small SNP panel for feed efficiency in beef cattle. *J. Anim. Sci.* 96, 375–397. doi: 10.1093/jas/sky020

Al-Jassim, R., and Sejian, V. (2015). Climate change and camel production: impact and contribution. *J. Camelid Sci.* 8, 1–17. Available online at: https://www.cabdirect.org/cabdirect/abstract/20163045120

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-terWengel, P., Pitt, D., et al. (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc. Natl. Acad. Sci. USA.* 113:6707. doi: 10.1073/pnas.1519508113

Almathen, F., Elbir, H., Bahbahani, H., Mwacharo, J., and Hanotte, O. (2018). Polymorphisms in MC1R and ASIP genes are associated with coat color variation in the Arabian camel. *J. Heredity* 109, 700–706. doi: 10.1093/jhered/esy024

Almutairi, S. E., Boujenane, I., Musaad, A., and Awad-Acharari, F. (2010). Genetic and nongenetic effects for milk yield and growth traits in Saudi camels. *Trop. Anim. Health Prod.* 42, 1845–1853. doi: 10.1007/s11250-010-9647-6

Al-Sobayil, K. A., Khalil, M. H., Al-Jobeile, H. S., Mohamed, K. M., and Salal, S. A. (2006). "Quantitative genetic analysis and evaluation for early growth performance in Saudi camels," in *International Scientific Conference on Camels* (Qassim), 201.

Al-Swailem, A. M.,Shehata, M. M., Abu-Duhier F. M, Al-Yamani EJ, Al-Busadah KA, Al-Arawi , M. S. et al. (2018). Sequencing, analysis, and annotation of expressed sequence tags for Camelus dromedaries. *PLoS ONE* 5:e0010720. doi: 10.1371/journal.pone.0010720

Ayadi, M., Aljumaah, R. S., Samara, E. M., Faye, B., and Caja, G. (2016). A proposal of linear assessment scheme for the udder of dairy camels (*Camelus dromedarius L.*). *Trop. Anim. Health Prod.* 48, 927–933. doi: 10.1007/s11250-016-1051-4

Ayadi, M., Musaad, A., Aljumaah, R. S., Samara, E. M., Abelrahman, M. M., Alshaikh, M. A., et al. (2013). Relationship between udder morphology traits, alveolar and cisternal milk compartments and machine milking performances of dairy camels (*Camelus dromedarius*). *J. Agric. Res.* 11, 790–797. doi: 10.5424/sjar/2013113-4060

Bornstein, S., and Younan, M. (2013). Significant veterinary research on the dromedary camels of Kenya: past and present. *J. Camelid Sci.* 6, 1–48. Available online at: https://www.cabdirect.org/cabdirect/FullTextPDF/2014/20143008728.pdf

Bower, M. A., McGivney, B. A., Campana, M. G., Gu, J., Andersson, L. S., Barrett, E., et al. (2012). The genetic origin and history of speed in the *Thoroughbred racehorse*. *Nature Commun.* 3:643. doi: 10.1038/ncomms1644

Burger, P. A., and Palmieri, N. (2014). Estimating the population mutation rate from a *de novo* assembled bactrian camel genome and cross-species comparison with dromedary ESTs. *J. Heredity* 105, 839–846. doi: 10.1093/jhered/est005

Caja, G., Diaz-Medina, E., Cabrera, S., Amann, O., Alama, O. H., El-Shafei, S., et al. (2013). "Comparison of traditional and modern systems for the individual identification of dromedary camels," in *ASAS-ADSA Joint Annual Meeting* Indianapolis, IN.

Chniter, M., Hammadi, M., Khorchani, T., Krit, R., Benwahada, A., and Hamouda, M. B. (2018). Classification of Maghrebi camels (*Camelus dromedarius*) according to theirtribal affiliation and body traits in southern Tunisia. *Emirates J. Food Agric.* 25, 625–634. doi: 10.9755/ejfa.v25i8.16096

Delgado, C. L., Rosegrant, M. W., Steinfeld, H., Ehui, S. K., and Courbois, C. (1999). *Livestock to 2020: The Next Food Revolution*. Washington, DC: IFPRI.

Dioli, M. (2016). Towards a rational camel breed judging: a proposed standard of a camel (*Camelus dromedarius*) milk breed. *J. Camel Pract. Res.* 23, 1–12. doi: 10.5958/2277-8934.2016.00001.1

Dittmann, M. T., Runge, U., Lang, R. A., Moser, D., Galeffi, C., Kreuzer, M., et al. (2014). Methane emission by camelids. *PLoS ONE* 9:e94363. doi: 10.1371/journal.pone.0094363

El-Bahrawy, K. A. (2018). Recent advances in dromedary camel reproduction: an Egyptian field experience. *Emirates J. Food Agric.* 27, 350–354. doi: 10.9755/ejfa.v27i4.19907

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

FAO. (2019). *R. Italy*. FAOSTAT.

Faye, B. (2011). Combating desertification: the added value of the camel farming. *Ann. Arid Zone* 50, 1–10. Available online at: http://publications.cirad.fr/une_notice.php?dk=568660

Faye, B. (2015). Role, distribution and perspective of camel breeding in the third millennium economies. *Emirates J. Food Agric.* 27, 318–327. doi: 10.9755/ejfa.v27i4.19906

Faye, B. (2018a). The enthusiasm for camel production. *Emirates J. Food Agric.* 30, 249–250. Available online at: https://www.ejfa.me/index.php/journal/article/view/1671

Faye, B. (2018b). The improvement of the camel reproduction performances: just a technical question? *Rev. Marocaine Sci. Agronom. Vétér.* 6, 265–269. Available online at: https://www.agrimaroc.org/index.php/Actes_IAVH2/article/view/607

Faye, B., Chaibou, M., and Gilles, V. (2012). Integrated impact of climate change and socioeconomic development on the evolutionof camel farming systems. *Br. J. Environ. Climate Change* 2, 227–244. doi: 10.9734/BJECC/2012/1548

Faye, B., Saint-Martin, G., Bonnet, P., Bengoumi, M., and Dia, M. L. (1997). *Guide de l'élevage du Dromadaire.* Libourne: Sanofi.

Gizaw, S., Getachew, T., Goshme, S., Valle-Zárate, A., van Arendonk, J. A., Kemp, S., et al. (2014). Efficiency of selection for body weight in a cooperative village breeding program of Menz sheep under smallholder farming system. *Animal* 8, 1249–1254. doi: 10.1017/S1751731113002024

Gonzalez-Recio, O., Coffey, M. P., and Pryce, J. E. (2014). On the value of the phenotypes in the genomic era. *J. Dairy Sci.* 97, 7905–7915. doi: 10.3168/jds.2014-8125

Gossner, C., Danielson, N., Gervelmeyer, A., Berthe, F., Faye, B., Kaasik Aaslav, K., et al. (2014). Human–dromedary camel interactions and the risk of acquiring zoonotic middle east respiratory syndrome coronavirus infection. *Zoonoses Public Health* 63, 1–9. doi: 10.1111/zph.12171

Hailu, Y., Hansen, E. B., Seifu, E., Eshetu, M., Ipsen, R., and Kappeler, S. (2016). Functional and technological properties of camel milk proteins: a review. *J. Dairy Res.* 83, 422–429. doi: 10.1017/S0022029916000686

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Hayes, B. J., Lewin, H. A., and Goddard, M. E. (2013). The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends Genet.* 29, 206–214. doi: 10.1016/j.tig.2012.11.009

Hemati, B., Banabazi, M., Shahkarami, S., Mohandesan, E., and Burger, P. (2017). Genetic diversity within bactrian camel population of Ardebil province. *Res. Anim. Prod.* 8, 197–202. doi: 10.29252/rap.8.16.192

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding.* Guelph: University of Guelph.

Hoffmann, I. (2010). Climate change and the characterization, breeding and conservation of animal genetic resources. *Anim. Genet.* 41, 32–46. doi: 10.1111/j.1365-2052.2010.02043.x

Jirimutu,Wang. Z., Ding, G., Chen, G., Sun, Y., et al. (2012). Genome sequences of wild and domestic bactrian camels. *Nat. Commun.* 3:1202. doi: 10.1038/ncomms2192

Kadim, I. T., Mahgoub, O., and Purchas, R. W. (2008). A review of the growth, and of the carcass and meat quality characteristics of the one-humped camel (*Camelus dromedaries*). *Meat Sci.* 80, 555–569. doi: 10.1016/j.meatsci.2008.02.010

Kess, T., Gross, J., Harper, F., and Boulding, E. G. (2016). Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkle *Littorina saxatilis*. *J. Molluscan Stud.* 82, 104–109. doi: 10.1093/mollus/eyv042

König , S., Simianer, H., and Willam, A. (2009). Economic evaluation of genomic breeding programs. *J. Dairy Sci.* 92, 382–391. doi: 10.3168/jds.2008-1310

Konuspayeva, G., Faye, B., and Loiseau, G. (2009). The composition of camel milk: a meta-analysis of the literature data. *J. Food Composit. Anal.* 22, 95–101. doi: 10.1016/j.jfca.2008.09.008

Manjunatha, B. M., Al-Bulushi, S., and Pratap, N. (2015). Synchronisation of the follicular wave with GnRH and PGF2α analogue for a timed breeding programme in dromedary camels (Camelus dromedarius). *Anim. Reprod. Sci.* 160, 23–29. doi: 10.1016/j.anireprosci.2015.06.023

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829. Available online at: http://www.genetics.org/content/157/4/1819.article-info

Mrode, R., Tarekegn, G. M., Mwacharo, J. M., and Djikeng, A. (2018). Invited review: genomic selection for small ruminants in developed countries: how applicable for the rest of the world? *Animal* 12, 1333–1340. doi: 10.1017/S1751731117003688

Mrode, R. A., Han, J., Mwacharo, J. M., and Koning, D. J. D. (2016). *Novel Tools to Inform Animal Breeding Programs.* Livestock and Fish Brief 14. Nairobi: ILRI.

Nikkhah, A. (2011). Equidae, camel, and yak milks as functional foods: a review. *J. Nutr. Food Sci.* 1:1. doi: 10.4172/2155-9600.1000116

Oulad-Belkhir, A., Chehma, A., and Faye, B. (2013). Phenotypic variability of two principal Algerian camel's populations (Targuiand Sahraoui). *Emirates J. Food Agric.* 25, 231–237. doi: 10.9755/ejfa.v25i3.15457

Raiymbek, G., Kadim, I., Konuspayeva, G., Mahgoub, O., Serikbayeva, A., and Faye, B. (2015). Discriminant amino-acid components of Bactrian (*Camelus bactrianus*) and Dromedary (*Camelus dromedarius*) meat. *J. Food Compos. Anal.* 41, 194–200. doi: 10.1016/j.jfca.2015.02.006

Ricard, A., Robert, C., Blouin, C., Baste, F., Torquet, G., Morgenthaler, C., et al. (2017). Endurance exercise ability in the horse: a trait with complex polygenic determinism. *Front. Genet.* 8:89. doi: 10.3389/fgene.2017.00089

Rodríguez-Ramilo, S. T., García-Cortés, L. A., and de Cara, M. Á.(2015). Artificial selection with traditional or genomic relationships: consequences in coancestry and genetic diversity. *Front. Genet.* 6:127. doi: 10.3389/fgene.2015.00127

Rothschild, M. F. (2004). Porcine genomics delivers new tools and results: this little piggy did more than just go to market. *Genet. Res.* 83, 1–6. doi: 10.1017/S0016672303006621

Rothschild, M. F., and Plastow, G. S. (2008). Impact of genomics on animal agriculture and opportunities for animal health. *Trends Biotechnol.* 26, 21–25. doi: 10.1016/j.tibtech.2007.10.001

Saalfeld, W. K., and Edwards, G. P. (2010). Distribution and abundance of the feral camel (*Camelus dromedarius*) in Australia. *Rangeland J.* 32, 1–9. doi: 10.1071/RJ09058

Skidmore, J. A., Morton, K. M., and Billah, M. (2013). Artificial insemination in dromedary camels. *Anim. Reprod. Sci.* 136, 178–186. doi: 10.1016/j.anireprosci.2012.10.008

Stangaferro, M. L., Wijma, R., Caixeta, L. S., Al-Abri, M. A., and Giordano, J. O. (2016). Use of rumination and activity monitoring for the identification of dairy cows with health disorders: part I. Metabolic and digestive disorders. *J. Dairy Sci.* 99, 7395–7410. doi: 10.3168/jds.2016-10907

Stiles, N. (1988). Le dromadaire contre l'avancée du désert. *La recherche* 20, 948–952.

Wurzinger, M., Sölkner, J., and Iñiguez, L. (2011). Important aspects and limitations in considering community-based breeding programs for low-input smallholder livestock systems. *Small Ruminant Res.* 98, 170–175. doi: 10.1016/j.smallrumres.2011.03.035

Zhou, H., Cheng, L., Azimu, W., Hodge, S., Edwards, G. R., and Hickford, J. G. H. (2015). Variation in the bovine FABP4 gene affects milk yield and milk protein content in dairy cows. *Sci. Rep.* 5:10023. doi: 10.1038/srep10023

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Abri and Faye. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 April 2019
doi: 10.3389/fgene.2019.00344

An Autosomal Translocation 73,XY,t(12;20)(q11;q11) in an Infertile Male Llama (*Lama glama*) With Teratozoospermia

*Malorie P. Baily[1], Felipe Avila[1], Pranab J. Das[2], Michelle A. Kutzler[3] and Terje Raudsepp[4]**

[1] School of Veterinary Medicine, University of California, Davis, Davis, CA, United States, [2] ICAR-National Research Centre on Pig, Assam, India, [3] Department of Animal and Rangeland Sciences, College of Agricultural Science, Oregon State University, Corvallis, OR, United States, [4] Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, United States

Structural chromosome abnormalities, such as translocations and inversions occasionally occur in all livestock species and are typically associated with reproductive and developmental disorders. Curiously, only a few structural chromosome aberrations have been reported in camelids, and most involved sex chromosomes. This can be attributed to a high diploid number (2n = 74) and complex chromosome morphology, which makes unambiguous identification of camelid chromosomes difficult. Additionally, molecular tools for camelid cytogenetics are sparse and have become available only recently. Here we present a case report about an infertile male llama with teratozoospermia and abnormal chromosome number 2n = 73,XY. This llama carries an autosomal translocation of chromosomes 12 and 20, which is the likely cause of defective spermatogenesis and infertility in this individual. Our analysis underlines the power of molecular cytogenetics methods over conventional banding-based chromosome analysis for explicit identification of normal and aberrant chromosomes in camelid karyotypes. This is the first case of a translocation and the first autosomal aberration reported in any camelid species. It is proof of principle that, like in other mammalian species, structural chromosome abnormalities contribute to reproductive disorders in camelids.

Keywords: camelids, cytogenetics, translocation, FISH, fertility, teratozoospermia

OPEN ACCESS

Edited by:
Pamela Burger, University of Veterinary Medicine, Austria

Reviewed by:
Rachele Antonacci, University of Bari Aldo Moro, Italy
Marek Switonski, Poznań University of Life Sciences, Poland

***Correspondence:**
Terje Raudsepp
traudsepp@cvm.tamu.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *21 October 2018*
Accepted: *29 March 2019*
Published: *16 April 2019*

Citation:
Baily MP, Avila F, Das PJ, Kutzler MA and Raudsepp T (2019) An Autosomal Translocation 73,XY,t(12;20)(q11;q11) in an Infertile Male Llama (Lama glama) With Teratozoospermia. Front. Genet. 10:344. doi: 10.3389/fgene.2019.00344

INTRODUCTION

Numerical and structural chromosome abnormalities are well-documented causes of congenital abnormalities and reproductive disorders in all livestock species (reviewed by Villagomez et al., 2009; Raudsepp and Chowdhary, 2016; Szczerbal and Switonski, 2016). Aberrations such as aneuploidies, deletions and duplications result in genetic overdose or haploinsufficiency, and may severely affect viability, development and/or reproduction. Translocations and inversions, on the other hand, are often balanced and do not cause loss or gain of the genetic material. Consequently, phenotypic effects of balanced rearrangements may not be so obvious regarding the viability

and appearance of the carrier. However, balanced structural rearrangements affect meiosis and gametogenesis, resulting in reduced fertility or infertility (Villagomez et al., 2009; Ghosh et al., 2016; Raudsepp and Chowdhary, 2016). Regardless, chromosome abnormalities are of concern in all livestock species and cytogenetic analysis is a routine approach for evaluating breeding animals and for testing animals with reproductive or developmental problems (Villagomez et al., 2009; Ducos et al., 2008; Lear and Bailey, 2008; Raudsepp and Chowdhary, 2016; Szczerbal and Switonski, 2016).

Compared to other domesticated species, clinical cytogenetics in alpacas, llamas and other camelids has progressed slowly. The first description of camelid karyotypes 50 years ago showed that all species have the same diploid number (2n = 74) with essentially similar chromosome morphology (Taylor et al., 1968). Since then, only a handful of reports have been published on chromosome aberrations in llamas and alpacas (reviewed by Raudsepp, 2014). These include only sex chromosome aneuploidies: two cases of X-monosomy (Hinrichs et al., 1997; Tibary, 2008), one case of X-trisomy (Tibary, 2008), two cases of XX female-to-male sex reversal (Wilker et al., 1994; Drew et al., 1999), and a dozen cases of XX/XY blood chimerism (Fowler, 1990; Hinrichs et al., 1997, 1999). So far, only two structural abnormalities have been described in camelids. One is the *Minute Chromosome Syndrome*, which has been found in infertile female alpacas and llamas (Drew et al., 1999; Tibary, 2008; Avila et al., 2014b; Fellows et al., 2014) and involves the smallest autosome, chromosome 36 (Avila et al., 2015). The second is an autosomal translocation in an infertile male llama that has been briefly and incompletely described in a study, whose main goal was the development of molecular cytogenetics tools for camelids (Avila et al., 2014b).

Reasons for the few clinical cytogenetics studies in camelids include a high chromosome number, a difficulty to identify chromosomes by conventional cytogenetic methods (Di Berardino et al., 2006; Avila et al., 2014b; Raudsepp, 2014), and the slow development of molecular cytogenetics tools for chromosome identification by fluorescence *in situ* hybridization (FISH) using DNA markers. While FISH has been a regular part of cytogenetics since the 1990s in most livestock/domestic species (Rubes et al., 2009), molecular cytogenetics tools for camelids became available only recently (Avila et al., 2014a,b, 2015).

Here, we revisit the prior partially studied case of the infertile male llama with an autosomal translocation (Avila et al., 2014b) and characterize it in detail clinically and cytogenetically using advanced semen imaging and conventional and molecular cytogenetic methods. This is the first autosomal translocation found in any camelid species.

MATERIALS AND METHODS

Ethics Statement

Procurement of blood and tissue samples followed the United States Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training. The protocols were approved by Institutional Animal Care and Use Committee as AUP #2009-115, AUP #2018-0342CA and CRRC#09-47 at Texas A&M University and ACUP #3817 at Oregon State University.

Case Description

A physically normal male llama born in 2000 was referred for andrological and cytogenetic evaluation due to infertility. At the time of referral, the animal was 3 years old and had never sired a cria, despite multiple breeding attempts to different fertile females.

Clinical Examination

Scrotum and testes were palpated for consistency to ensure that no gross pathology was present in the external genitalia. Testicular size was measured mechanically with sliding calipers. Testes and accessory glands were imaged with a real-time ultrasound scanner using a 5 MHz probe (Sonovet SV600, Universal Medical Systems Inc.) and testicular blood flow was measured in the marginal (MA) and supratesticular arteries (TA) by Doppler ultrasonography with an L12-5 probe (Kutzler et al., 2011).

Semen Analysis

Semen was collected for six days using a ruminant artificial vagina inserted into a custom-designed llama phantom. Slides were prepared from each ejaculate using eosin-nigrosin staining method (Murcia-Robayo et al., 2018). Sperm morphology was first examined under a light microscope at 1000 x magnification, followed by transmission electron microscopy (TEM) with FEI TITAN 80–200 TEM/STEM with Chem-iSTEM Technology following standard procedures (Alvarenga et al., 2000; Pesch et al., 2006).

Cell Cultures, Chromosome Preparation and Karyotyping

Samples for karyotyping included peripheral blood in Na-heparin (Becton Dickinson) as well as an ear clip in sterile Hank's balanced salt solution containing 1X Antibiotic-Antimycotic solution (Gibco). The ear clip was used to establish primary fibroblast cultures. Metaphase chromosome preparations were obtained from short-term blood lymphocyte cultures or skin fibroblast cultures, according to standard procedures (Raudsepp and Chowdhary, 2008; Avila et al., 2014b). Chromosomes were stained with Giemsa for initial counting. Refined chromosome analysis and karyotyping were carried out by GTG banding (Seabright, 1971). Images for 20 cells were captured for each technique using an Axioplan2 microscope (Carl Zeiss) and IKAROS (MetaSystems GmbH) software. Twenty cells were karyotyped and chromosomes were arranged into karyograms following the nomenclature proposed by Balmus et al. (2007) and adopted for the alpaca by Avila et al. (2014b).

DNA Isolation and Analysis of Sex Chromosomes by PCR

Peripheral blood was collected in EDTA vacutainers (Becton Dickinson) and DNA was isolated with Gentra Puregene Blood Kit (Qiagen), following the manufacturer's protocol. Genomic DNA was used as a template for PCR reactions with alpaca primers for the Y-linked *SRY* gene (F:5′-GTCAAGCGCCCC ATGAATGC-3′; R: 5′- CGTAGTCTCTGTGCCTCCTC-3′; 170 bp) (Drew et al., 1999) and the X-linked androgen receptor (*AR*) gene (F: 5′- GCTTTCCAGAACCTGTTCCA -3′; R: 5′-GCCTCTGCTCTGGACTTGTG -3′; 204 bp).

Fluorescence *in situ* Hybridization (FISH)

Painting probes generated from flow-sorted alpaca X and Y chromosomes (Avila et al., 2014b) were used to test the presence and integrity of sex chromosomes. The origin of the autosomal translocation was investigated through series of dual-color FISH experiments with chromosome-specific markers (BAC clones from CHORI-246 BAC library[1] derived from the alpaca whole genome cytogenetic map (Avila et al., 2014a). BAC DNA isolation, labeling and FISH were performed following standard protocols (Raudsepp and Chowdhary, 2008; Avila et al., 2014b). Images for a minimum of 10 metaphase spreads were captured for each experiment and analyzed with a Zeiss Axioplan2 fluorescence microscope equipped with Isis Version 5.2 (MetaSystems GmbH) software.

RESULTS

Clinical and Semen Analysis

On physical examination, no general or reproductive abnormalities were found in the male llama. The testicles were of normal palpable consistency and size (**Table 1**). Ultrasonographic imaging of the testes and accessory sex glands revealed no abnormalities, with both the prostate and bulbourethral gland showing a typical homogeneous echotexture. Testicular blood flow measurements were within normal range for camelids (**Table 2**; Kutzler et al., 2011).

However, analysis of semen samples from six consecutive days, revealed that the percentage of morphologically normal sperm (**Figure 1A**) was low and ranged from 31.3 to 40.5% (**Table 3**). The most common abnormality observed was an abnormally thickened midpiece (**Table 3** and **Figure 1B**), which occurred in 10.8–20.7% of the sperm evaluated. The occurrence of nuclear vacuolation was also high, ranging from 3.3 to 13.7% (**Table 3** and **Figure 1B**). The semen abnormalities that were observed by light microscopic examination were confirmed and refined by TEM analysis. The latter showed the presence of both acrosomal (**Figure 2A**) and nuclear vacuolations

[1]https://bacpacresources.org/

TABLE 1 | Testicular measurements.

Testis	Length, cm	Width, cm	Diameter, cm
Left	4.2	2.2	2.4
Right	4.5	2.0	2.5

TABLE 2 | Testicular blood flow measurements by Doppler ultrasonography of the marginal (MA) and supratesticular arteries (TA).

Testis	MA PSV, cm/s	MA EDV, cm/s	MA RI	TA PSV, cm/s	TA EDV, cm/s	TA RI
Left	6.30	4.80	0.24	11.10	3.40	0.69
Right	9.30	7.80	0.16	15.80	5.40	0.66
Average	7.80	6.30	0.20	13.45	4.40	0.68

PSV, peak systolic volume; EDV, end diastolic volume; RI, resistance index.

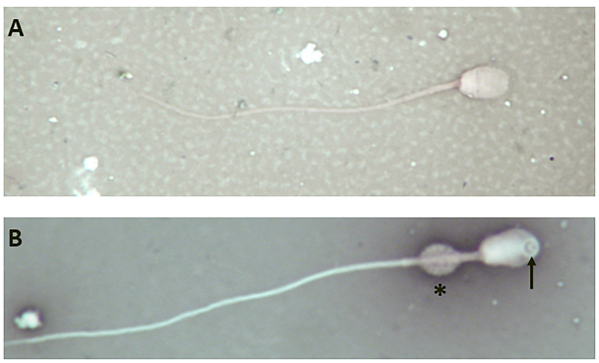


FIGURE 1 | Light microscope images of eosin-nigrosine stained llama sperm. **(A)** Normal spermatozoon. **(B)** Abnormal spermatozoon showing a large nuclear vacuole (arrow) and an abnormally thickened midpiece (asterisk); magnification 1000×.

TABLE 3 | Sperm morphology analysis results from daily semen collection.

Sperm Characteristic	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Normal, %	40.5	31.3	35	33	36	34.9
Abnormally thickened midpiece, %	19.8	19.5	10.8	20.5	20.7	18.7
Detached head, %	6.6	15.2	16	10.2	13.7	9.7
Abnormally thick tail, %	0.82	0	0	0	0	0
Severely coiled tail, %	9.9	9.3	5.4	4.2	4.3	4
Abnormally long, skinny head, %	1.6	5.1	0	0	2.5	2.4
Short fat head, %	2.5	1.7	0	1.7	3.4	1.6
Bent midpiece, %	0	3.4	8.1	0	0.86	0
Severely bent midpiece, %	2.5	3.4	10.8	5.1	5.17	10.5
Microcephalia, %	9	2.5	5.4	4.2	6.8	7.3
Pyriform head, %	0.8	2.5	2.7	1.7	0.9	0.2
Broken midpiece, %	1.6	0.8	0	2.6	0	2.4
Nuclear vacuoles, %	3.3	3.3	5.4	13.7	5.19	4
Severely bent tail, %	0	0	0	0	0	1.6
Bent tail, %	0	0	0	0.8	0.86	0
Bent neck, %	0.8	1.7	0	0.8	1.7	0
Broken Neck, %	0	0	0	0	0	1.6
Total sperm counted	121	118	37	117	116	123

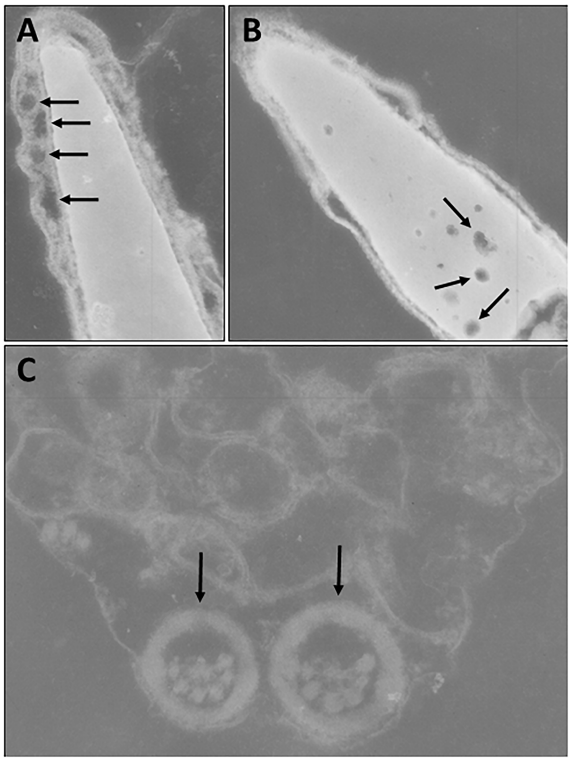


FIGURE 2 | Transmission electron microscopy of sperm. **(A)** Abnormal spermatozoon with acrosomal vacuolation (arrows). **(B)** Abnormal spermatozoon with nuclear vacuolation (arrows). **(C)** Abnormal axoneme formation (arrows).

(**Figure 2B**) as well as several sperm with abnormal axoneme formation (**Figure 2C**). Based on these findings, we concluded that the phenotypically normal-looking llama (**Figure 3A**) has severe teratozoospermia.

Molecular Cytogenetic Analysis

Analysis of genomic DNA by PCR showed that the llama was positive for both the *SRY* and the *AR* genes, the expected profile of normal males.

Initial karyotyping of Giemsa-stained chromosomes indicated an abnormal diploid number of 2n = 73 in all metaphase spreads studied. This was confirmed via refined analysis by GTG banding, which also revealed the presence of a large submetacentric derivative chromosome (**Figures 3B,C**) that resembled the X chromosome in size, morphology, and banding pattern (**Figure 4A**). However, FISH analysis with alpaca X and Y chromosome painting probes showed that all cells contained only one X and one Y chromosome (**Figure 4B**), thus ruling out a sex-linked origin for the derivative chromosome. This suggested that the derivative chromosome was a result of a translocation of two medium-sized autosomes, which also explained the abnormal chromosome number of 73. Attempts to identify the autosomes involved in translocation by GTG banding were inconclusive due to morphological and banding similarities amongst different chromosome pairs in camelids (**Figure 3C**; Balmus et al., 2007; Avila et al., 2014b). Therefore, we conducted multiple dual-color FISH experiments with pairs of alpaca BAC clones containing markers specific to likely candidate chromosomes for the aberration. This analysis revealed that the derivative chromosome was the result of a translocation between chr12 and chr20 (**Figures 5A,B**). The markers that identified the derivative chromosome were BACs 4J13 and 154O16 for chr12 and BAC 92P17 (**Figure 5C**) for chr20 (Avila et al., 2014a). The karyotype of the infertile male llama was denoted as 73,XY,t(12;20).

DISCUSSION

The case of an autosomal translocation in an infertile male llama described herein is the first report of a translocation in any camelid species.

While FISH results demonstrated that the derivative chromosome harbors the long arms (q-arms) of chr12 and chr20 (**Figure 5**), we could not trace the location of the short arms (p-arms) of these chromosomes because no DNA markers have been, as yet, mapped to these regions (Avila et al., 2014a). On the other hand, if the fusion occurred between chr12p and chr20p (**Figure 5C**), additional rearrangements had to have taken place to define the centromere of the derivative chromosome. Thus, we consider the fusion of short arms unlikely. It is more plausible that the derivative chromosome resulted from a centric fusion of chr12q and chr20q, with subsequent loss of 12p and 20p. This is in keeping with observations that the short arms of most camelid chromosomes are heterochromatic and vary in size between individuals, as well as between homologs as shown by C-banding (Bianchi et al., 1986; Di Berardino et al., 2006; Balmus et al., 2007; Avila et al., 2014b). The fact that gene sequences have been assigned to the long arms of all 36 alpaca autosomes, but only to short arms of 6 of these (Avila et al., 2014a), further suggests their heterochromatic nature. It must be noted that because C-banding does not allow chromosome identification in camelids, it was not possible to evaluate heterochromatin in chr12p and 20p by this method. Attempts to combine C-banding with FISH for chromosome identification were also unsuccessful. However, heterochromatin at camelid centromeres and chromosome arms is AT-rich and stains bright with DAPI (black with inverted DAPI), which is the case with chr12p and chr20p (**Figure 5B**). Therefore, we theorize that the translocation did not cause loss of functionally important genetic material in somatic cells and can be considered as a balanced translocation. This is consistent with the overall normal appearance and viability of the carrier llama (**Figure 3A**).

Balanced translocations typically disturb meiotic pairing and segregation, resulting in the production of both genetically balanced and unbalanced gametes (Ghosh et al., 2016; Raudsepp and Chowdhary, 2016). The latter, if involved in fertilization, will cause embryonic death and, thus, subfertility of the translocation carrier. Such cases have been abundantly described in all livestock species (Ghosh et al., 2016; Raudsepp and Chowdhary, 2016; Szczerbal and Switonski, 2016). The llama in the present study has a more pronounced abnormal reproductive phenotype (teratozoospermia and sterility) than cases previously described in other livestock species, suggesting that the translocation may have disrupted function of genes important for normal

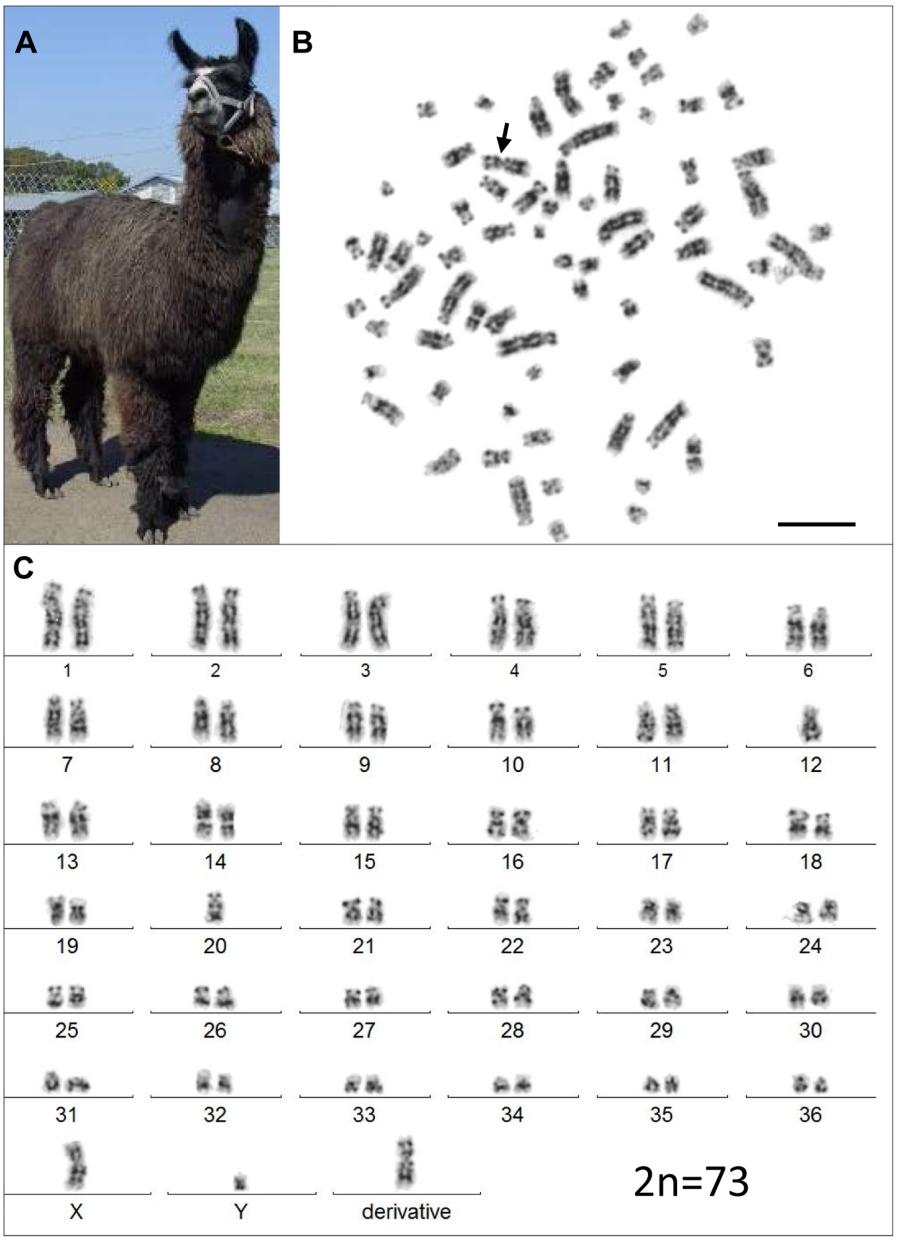


FIGURE 3 | The infertile male llama and his karyotype. **(A)** The male llama showing overall normal appearance and viability. **(B)** A GTG-banded metaphase spread with the derivative chromosome (arrow). **(C)** Its corresponding karyotype (2n = 73). Bar 10 μm.

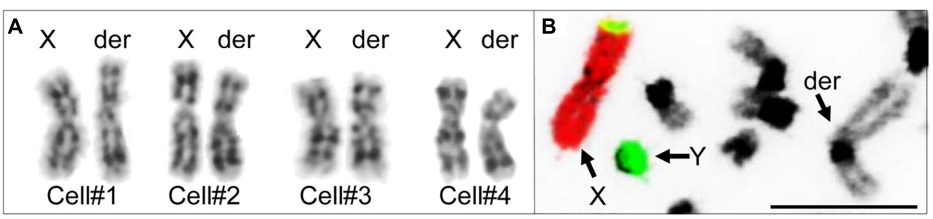


FIGURE 4 | Investigating the origin of the derivative chromosome. **(A)** Comparison of GTG-banded chrX and the derivative chromosome in four different cells. **(B)** FISH with alpaca chrX (red) and chrY (green) painting probes showing non-sex chromosomal origin of the derivative chromosome (modified from Avila et al., 2014b). Bar 10 μm.

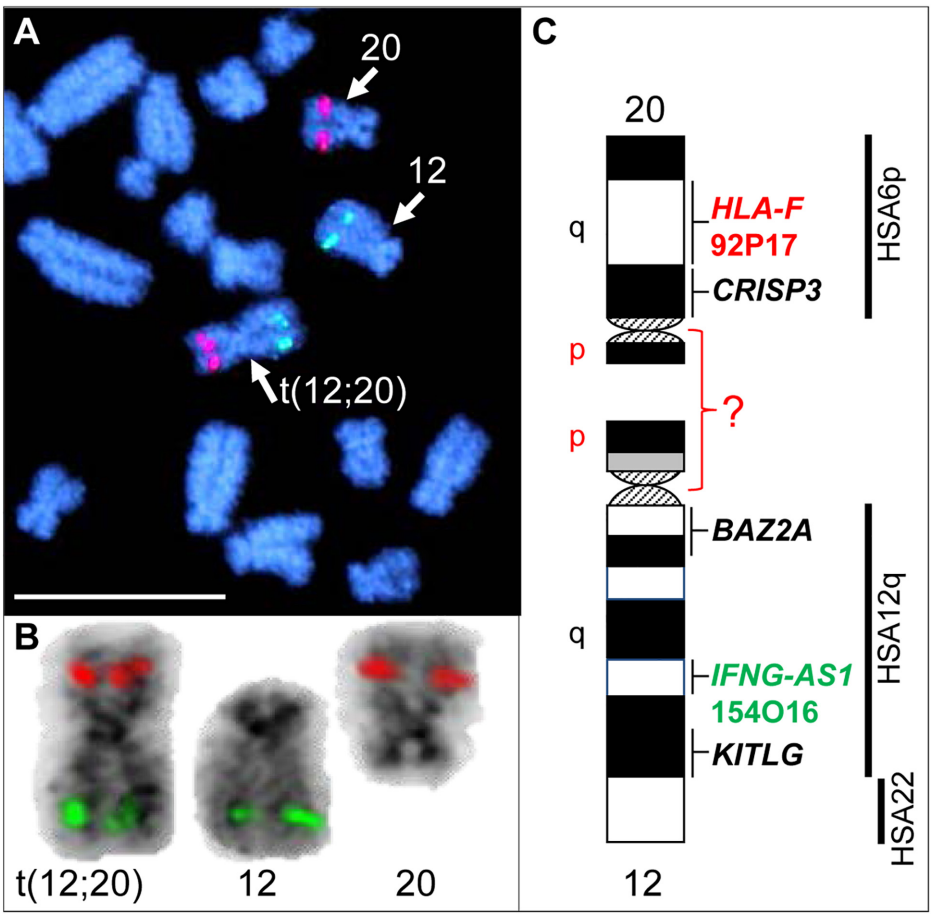


FIGURE 5 | FISH with chr12 (green; BAC 154O16; *IFNG-AS1*) and chr20 (red; BAC 92P17; *HLA-F*) specific probes identifies the origin of the derivative chromosome. **(A)** Partial metaphase showing the location of normal chr12, normal chr20 and their translocation product. **(B)** Derivative chromosome, normal chr12 and normal chr20 (inverted) with corresponding FISH signals. **(C)** Schematic G-banded ideograms of chr12 and chr20 with information about mapped genes and human (HSA) homology (far right). Bar 10 μm.

spermatogenesis and/or fertilization. For example, camelid chromosome 20 is homologous to human (HSA) chromosome 6p and harbors the major histocompatibility complex (MHC) and a cluster of genes encoding for cysteine rich secretory proteins – the *CRISP* genes (Avila et al., 2014a). Of these, *CRISP3* was cytogenetically mapped very close to the presumed translocation break/fusion point (**Figure 5C**). CRISP proteins are expressed in the male reproductive tract and have known roles in sperm function, sperm-egg interactions, and overall fertility in many mammalian species, including camelids (Waheed et al., 2015; Gottschalk et al., 2016; Fedorka et al., 2017). The counterpart of the translocation, chr12, is homologous to HSA12q and part of HSA22 (Balmus et al., 2007; Avila et al., 2014a; **Figure 5C**), but no male fertility genes have been mapped to this camelid chromosome. Therefore, even though the involvement of *CRISP3* in the translocation is an appealing target for speculation, the molecular consequences of this rearrangement remain unknown and will be particularly interesting for follow-up.

In summary, we described the first chromosomal translocation in camelids, its likely causative relationship with teratozoospermia and infertility, and demonstrated the power of molecular cytogenetic approaches for the detection of structural aberrations in species with complex karyotypes. Characterization of molecular and functional consequences of such rearrangements, however, requires further research and improved knowledge about camelid genomes.

AUTHOR CONTRIBUTIONS

MK and TR initiated and designed the study. MB, FA, MK, PD, and TR conducted experimental work and data analysis. MB, FA, and TR wrote the manuscript with input from all authors.

FUNDING

This study was supported by grants from Alpaca Research Foundation 2009–2011, Morris Animal Foundation D09LA-004, and Willamette Valley Llama Association.

REFERENCES

Alvarenga, M. A., Landim-Alvarenga, F. C., Moreira, R. M., and Cesarino, M. M. (2000). Acrosomal ultrastructure of stallion spermatozoa cryopreserved with ethylene glycol using two packaging systems. *Equine Vet. J.* 32, 541–545. doi: 10.2746/042516400777584749

Avila, F., Baily, M. P., Merriwether, D. A., Trifonov, V. A., Rubes, J., Kutzler, M. A., et al. (2015). A cytogenetic and comparative map of camelid chromosome 36 and the minute in alpacas. *Chromosome Res.* 23, 237–251. doi: 10.1007/s10577-014-9463-3

Avila, F., Baily, M. P., Perelman, P., Das, P. J., Pontius, J., Chowdhary, R., et al. (2014a). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*lama pacos*). *Cytogenet. Genome. Res.* 144, 196–207. doi: 10.1159/000370329

Avila, F., Das, P. J., Kutzler, M., Owens, E., Perelman, P., Rubes, J., et al. (2014b). Development and application of camelid molecular cytogenetic tools. *J. Hered.* 105, 858–869. doi: 10.1093/jhered/ess067

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C., Alkalaeva, E. S., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative cetartiodactyla ancestral karyotype. *Chromosome Res.* 15, 499–515. doi: 10.1007/s10577-007-1154-x

Bianchi, N. O., Larramendy, M. L., Bianchi, M. S., and Cortes, L. (1986). Karyological conservation in south american camelids. *Experientia* 42, 622–624. doi: 10.1007/BF01955563

Di Berardino, D., Nicodemo, D., Coppola, G., King, A. W., Ramunno, L., Cosenza, G. F., et al. (2006). Cytogenetic characterization of alpaca (*Lama pacos*, fam. Camelidae) prometaphase chromosomes. *Cytogenet. Genome Res.* 115, 138–144. doi: 10.1159/000095234

Drew, M. L., Meyers-Wallen, V. N., Acland, G. M., Guyer, C. L., and Steinheimer, D. N. (1999). Presumptive sry-negative xx sex reversal in a llama with multiple congenital anomalies. *J. Am. Vet. Med. Assoc.* 215, 1134–1139.

Ducos, A., Revay, T., Kovacs, A., Hidas, A., Pinton, A., Bonnet-Garnier, A., et al. (2008). Cytogenetic screening of livestock populations in europe: an overview. *Cytogenet. Genome. Res.* 120, 26–41. doi: 10.1159/000118738

Fedorka, C. E., Scoggin, K. E., Squires, E. L., Ball, B. A., and Troedsson, M. H. T. (2017). Expression and localization of cysteine-rich secretory protein-3 (crisp-3) in the prepubertal and postpubertal male horse. *Theriogenology* 87, 187–192. doi: 10.1016/j.theriogenology.2016.08.027

Fellows, E., Kutzler, M., Avila, F., Das, P. J., and Raudsepp, T. (2014). Ovarian dysgenesis in an alpaca with a minute chromosome 36. *J. Hered.* 105, 870–874. doi: 10.1093/jhered/ess069

Fowler, M. (1990). Twinning in llamas. *Int Camelid J.* 4, 35–38.

Ghosh, S., Das, P. J., Avila, F., Thwaits, B. K., Chowdhary, B. P., and Raudsepp, T. (2016). A non-reciprocal autosomal translocation 64,xx, t(4;10)(q21;p15) in an arabian mare with repeated early embryonic loss. *Reprod. Domest. Anim.* 51, 171–174. doi: 10.1111/rda.12636

Gottschalk, M., Metzger, J., Martinsson, G., Sieme, H., and Distl, O. (2016). Genome-wide association study for semen quality traits in german warmblood stallions. *Anim. Reprod. Sci.* 171, 81–86. doi: 10.1016/j.anireprosci.2016.06.002

Hinrichs, K., Buoen, L. C., and Ruth, G. R. (1999). Xx/xy chimerism and freemartinism in a female llama co-twin to a male. *J. Am. Vet. Med. Assoc.* 215, 1140–1141.

Hinrichs, K., Horin, S. E., Buoen, L. C., Zhang, T. Q., and Ruth, G. R. (1997). X-chromosome monosomy in an infertile female llama. *J. Am. Vet. Med. Assoc.* 210, 1503–1504.

Kutzler, M., Tyson, R., Grimes, M., and Timm, K. (2011). Determination of testicular blood flow in camelids using vascular casting and color pulsed-wave doppler ultrasonography. *Vet. Med. Int.* 2011:638602. doi: 10.4061/2011/638602

Lear, T. L., and Bailey, E. (2008). Equine clinical cytogenetics: the past and future. *Cytogenet. Genome. Res.* 120, 42–49. doi: 10.1159/000118739

Murcia-Robayo, R. Y., Jouanisson, E., Beauchamp, G., and Diaw, M. (2018). Effects of staining method and clinician experience on the evaluation of stallion sperm morphology. *Anim. Reprod. Sci.* 188, 165–169. doi: 10.1016/j.anireprosci.2017.11.021

Pesch, S., Bostedt, H., Failing, K., and Bergmann, M. (2006). Advanced fertility diagnosis in stallion semen using transmission electron microscopy. *Anim. Reprod. Sci.* 91, 285–298. doi: 10.1016/j.anireprosci.2005.04.004

Raudsepp, T. (2014). "Cytogenetics and infertility, in Chris Cebra," in *Llama and Alpaca Care: Medicine, Surgery, Reproduction, Nutrition and Herd Health*, eds D. E. Anderson, A. Tibary, R. J. Van Saun, and L. Johnson (Philadelphia, PA: Elsevier Inc.), 243–249. doi: 10.1016/B978-1-4377-2352-6.00021-3

Raudsepp, T., and Chowdhary, B. P. (2008). Fish for mapping single copy genes. *Methods Mol. Biol.* 422, 31–49. doi: 10.1007/978-1-59745-581-7_3

Raudsepp, T., and Chowdhary, B. P. (2016). Chromosome aberrations and fertility disorders in domestic animals. *Annu. Rev Anim. Biosci.* 4, 15–43. doi: 10.1146/annurev-animal-021815-111239

Rubes, J., Pinton, A., Bonnet-Garnier, A., Fillon, V., Musilova, P., Michalova, K., et al. (2009). Fluorescence in situ hybridization applied to domestic animal cytogenetics. *Cytogenet. Genome Res.* 126, 34–48. doi: 10.1159/000245905

Seabright, M. (1971). A rapid banding technique for human chromosomes. *Lancet* 2, 971–972. doi: 10.1016/S0140-6736(71)90287-X

Szczerbal, I., and Switonski, M. (2016). "Chromosome abnormalities in domestic animals as causes of disorders of sex development or impaired fertility," in *Insights from Animal Reproduction*, ed. R. P. Carreira (London: INTECH), 207–225.

Taylor, K. M., Hungerford, D. A., Snyder, R. L., and Ulmer, F. A. Jr. (1968). Uniformity of kryotypes in the camelidae. *Cytogenetics* 7, 8–15. doi: 10.1159/000129967

Tibary, A. (2008). "Reproductive disorders in alpacas and llamas," in *Proceedings of the 1st International Workshop on Camelid Genetics*, (Scottsdale, AZ), 22–24.

Villagomez, D. A., Parma, P., Radi, O., Di Meo, G., Pinton, A., Iannuzzi, L., et al. (2009). Classical and molecular cytogenetics of disorders of sex development in domestic animals. *Cytogenet. Genome Res.* 126, 110–131. doi: 10.1159/000245911

Waheed, M. M., Ghoneim, I. M., and Alhaider, A. K. (2015). Seminal plasma and serum fertility biomarkers in dromedary camels (*camelus dromedarius*). *Theriogenology* 83, 650–654. doi: 10.1016/j.theriogenology.2014.10.033

Wilker, C. E., Meyers-Wallen, V. N., Schlafer, D. H., Dykes, N. L., Kovacs, A., and Ball, B. A. (1994). Xx sex reversal in a llama. *J. Am. Vet. Med. Assoc.* 204, 112–115.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Baily, Avila, Das, Kutzler and Raudsepp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 April 2019
doi: 10.3389/fgene.2019.00340

Comparative FISH-Mapping of *MC1R*, *ASIP*, and *TYRP1* in New and Old World Camelids and Association Analysis With Coat Color Phenotypes in the Dromedary (*Camelus dromedarius*)

*Fahad Alshanbari[1], Caitlin Castaneda[1], Rytis Juras[1], Andrew Hillhouse[2], Mayra N. Mendoza[3], Gustavo A. Gutiérrez[3], Federico Abel Ponce de León[4] and Terje Raudsepp[1]**

[1] Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX, United States, [2] Institute for Genome Sciences and Society, Texas A&M University, College Station, TX, United States, [3] Animal Breeding Program, National Agrarian University La Molina, Lima, Peru, [4] Department of Animal Science, University of Minnesota, Minneapolis, MN, United States

OPEN ACCESS

Edited by:
Pamela Burger, University of Veterinary Medicine, Austria

Reviewed by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom
Martin Plášil, University of Veterinary and Pharmaceutical Sciences Brno, Czechia

***Correspondence:**
Terje Raudsepp
traudsepp@cvm.tamu.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 December 2018*
Accepted: *29 March 2019*
Published: *16 April 2019*

Citation:
Alshanbari F, Castaneda C, Juras R, Hillhouse A, Mendoza MN, Gutiérrez GA, Ponce de León FA and Raudsepp T (2019) Comparative FISH-Mapping of MC1R, ASIP, and TYRP1 in New and Old World Camelids and Association Analysis With Coat Color Phenotypes in the Dromedary (Camelus dromedarius). Front. Genet. 10:340. doi: 10.3389/fgene.2019.00340

Melanocortin 1 receptor (*MC1R*), the agouti signaling protein (*ASIP*), and tyrosinase related protein 1 (*TYRP1*) are among the major regulators of pigmentation in mammals. Recently, *MC1R* and *ASIP* sequence variants were associated with white and black/dark brown coat colors, respectively, in the dromedary. Here we confirmed this association by independent sequencing and mutation discovery of *MC1R* and *ASIP* coding regions and by TaqMan genotyping in 188 dromedaries from Saudi Arabia and United States, including 38 black, 53 white, and 97 beige/brown/red animals. We showed that heterozygosity for a missense mutation c.901C > T in *MC1R* is sufficient for the white coat color suggesting a possible dominant negative effect. Likewise, we confirmed that the majority of black dromedaries were homozygous for a frameshift mutation in *ASIP* exon 2, except for 4 animals, which were heterozygous. In search for additional mutations underlying the black color, we identified another frameshift mutation in *ASIP* exon 4 and 6 new variants in *MC1R* including a significantly associated SNP in 3′UTR. In pursuit of sequence variants that may modify dromedary wild-type color from dark-reddish brown to light beige, we identified 4 SNPs and one insertion in *TYRP1* non-coding regions. However, none of these were associated with variations in wild-type colors. Finally, the three genes were cytogenetically mapped in New World (alpaca) and Old World (dromedary and Bactrian camel) camelids. The *MC1R* was assigned to chr21, *ASIP* to chr19 and *TYRP1* to chr4 in all 3 species confirming extensive conservation of camelid karyotypes. Notably, while the locations of *ASIP* and *TYRP1* were in agreement with human-camelid comparative map, mapping *MC1R* identified

a new evolutionary conserved synteny segment between camelid chromosome 21 and HSA16. The findings contribute to coat color genomics and the development of molecular tests in camelids and toward the chromosome level reference assemblies of camelid genomes.

Keywords: camelids, ASIP, MC1R, TYRP1, FISH, TaqMan assay

INTRODUCTION

Mammalian coat color is a phenotypic trait that serves for camouflage and communication in the wild, and has been a target for selection by humans in farm and companion species since their domestication (Andersson, 2001; Cieslak et al., 2011). As a result, domestic animals display a perplexing variety of colors, patterns and markings, which reflect the genetic diversity of a breed or species, as well as historic and aesthetic preferences or commercial needs of humans.

Many genes regulate coat color. This was already noted by Haldane (1927) over 90 years ago when he studied color genetics in rodents and carnivores and suggested that there are at least 20 different color genes in mammals. Since then, approximately 150 coat-color associated genes have been described in mice, humans, and domestic animals (Schmutz and Berryere, 2007; Bellone, 2010; Cieslak et al., 2011; Reissmann and Ludwig, 2013), whereas *Color Genes* database[1] lists 378 mouse loci with their human and zebrafish homologs that are associated with various pigmentation phenotypes.

Despite the large number of genes involved, the production, amount and distribution of main pigments, the brown/black eumelanin and the red/yellow pheomelanin, are controlled by just a few major pigmentation genes (Rees, 2003; Bellone, 2010; Reissmann and Ludwig, 2013; Suzuki, 2013). These include melanocortin 1 receptor (*MC1R*), agouti signaling protein (*ASIP*) and tyrosinase related protein 1 (*TYRP1*). Melanocortin 1 receptor is the key switch between the synthesis of eumelanin or pheomelanin; ASIP is an antagonist ligand that regulates MC1R signaling by inhibiting the MC1R receptor, and TYRP1 is a melanogenic enzyme that influences the quantity and quality of melanins (Pielberg, 2004; Bellone, 2010; Sturm and Duffy, 2012; Suzuki, 2013). Associations between basic coat colors and DNA sequence polymorphisms in *MC1R, ASIP, TYRP1*, are known for most domestic species (Rieder et al., 2001; Pielberg, 2004; Schmutz and Berryere, 2007; Bellone, 2010; Cieslak et al., 2011), and are routinely used for genetic testing.

In contrast to other domestic species, coat color genomics in camelids had a late start, even though fiber color is an important trait for the alpaca industry (Morante et al., 2009) and there is an interest for breeding white or black dromedaries in some Arabian countries (Almathen et al., 2018). A few studies in alpacas have associated mutations in *ASIP* with the black color (Feeley et al., 2011; Chandramohan et al., 2013) and identified *MC1R* mutations that may determine light phenotypes, though the findings about the alpaca *MC1R* remain inconclusive (Feeley and Munyard, 2009; Guridi et al., 2011; Chandramohan et al., 2015). Research on dromedary color genes is even more recent with just two publications. The first study revealed that a frameshift mutation in the *KIT* gene explains some, though not all forms of white-spotting phenotypes in the dromedary (Holl et al., 2017). The most recent study identified a missense mutation in *MC1R* that is associated with the white color, and a deletion and a single nucleotide polymorphism (SNP) in *ASIP* exon 2 that are associated with the black/dark brown color in dromedaries (Almathen et al., 2018). Current reference genomes for the alpaca and the dromedary are in scaffolds and not assigned to chromosomes[2]. Because of this, chromosomal location is known only for the few coat color genes that were included in the alpaca whole genome cytogenetic map (Avila et al., 2014a). Among the main pigmentation genes, *ASIP* and *TYRP1* have been mapped in the alpaca but not in other camelids, whereas *MC1R* is not mapped in any camelid species.

The aim of this study is to confirm and refine the recently reported *MC1R* and *ASIP* mutations for white and black coat color in dromedaries, and search for novel color-related variants in *TYRP1*. We compare the accuracy of genotyping the white and black mutations in large dromedary populations by direct sequencing and with a TaqMan™ assay. Finally, we cytogenetically map *ASIP*, *MC1R,* and *TYRP1* in three camelid species.

MATERIALS AND METHODS

Ethics Statement

Procurement of peripheral blood was performed according to the United States Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training. These protocols were approved by Animal Use Protocol AUP #2011-96, # 2018-0342 CA and CRRC #09-47 at Texas A&M University.

Animals and Phenotypes

We sampled 188 dromedaries originating from Saudi Arabia (SA; n = 171) and from the United States (US; n = 17). Coat color phenotypes were determined by visual inspection, recorded in written notes and/or photos, and were as follows: white/cream (n = 53), black/dark brown (n = 38), and brown/beige (n = 97) (**Figure 1**). Two brown dromedaries had white markings and blue eyes. We use '*brown*' as a generic term to denote animals with wild-type coat color, which can range from light beige to darker reddish–brown with either matching or darker tail and hump.

[1]http://www.espcr.org/micemut/

[2]https://www.ncbi.nlm.nih.gov/genome

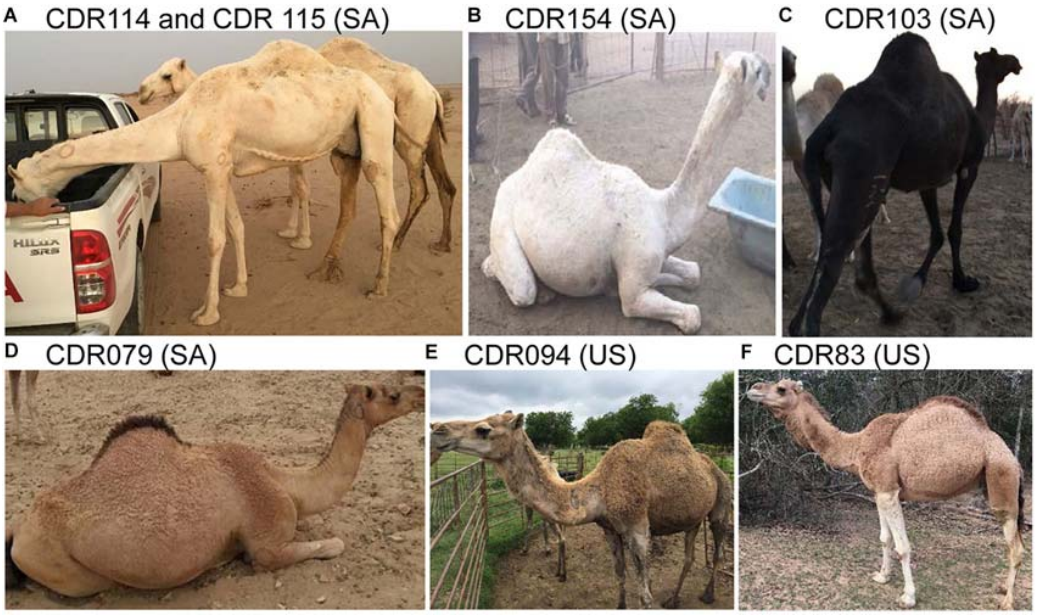


FIGURE 1 | Examples of animals and coat colors used for this study. **(A)** White/cream; **(B)** White; **(C)** Black; **(D)** Reddish brown with dark hump and tail; **(E)** Medium brown; and **(F)** Light brown; SA, Saudi Arabia; US, United States; dromedary IDs and genotypes are in **Supplementary Table S1**.

Supplementary Table S1 presents summary information for all animals and phenotypes.

Samples

Blood was collected by jugular venipuncture into EDTA-containing Vacutainers (Becton Dickinson).

DNA Isolation

Genomic DNA was isolated from peripheral blood lymphocytes using Gentra Puregene Blood Kit (Qiagen) following the manufacturer's protocol, or by standard phenol-chloroform method (Sambrook et al., 1989). We evaluated DNA quality and quantity by NanoDrop 2000 spectrophotometer (Thermo Scientific) and by 1% agarose gel electrophoresis.

Primers, PCR and Sequencing

We used the available sequence information for the dromedary and alpaca *MC1R, ASIP,* and *TYRP1* in NCBI[3], UCSC[4], and Ensembl[5] genome browsers, or sequences of the Bactrian camel (Wu et al., 2014) and Primer3 software (Untergasser et al., 2012) to design primers. For *ASIP* and *TYRP1*, primers were designed to amplify all exons and exon–intron boundaries. For *MC1R*, primers were designed to amplify overlapping fragments covering the single exon and the 5′UTR. Primer details are presented in **Table 1**. PCR was conducted in 10 μL reactions containing 50 ng dromedary genomic DNA and 0.5 unit of JumpStart Taq ReadyMix (Sigma Aldrich). For *MC1R,* primers 5′UTR.1 F and 5′UTR.2 R (**Table 1**) were combined to amplify the entire 2 kb of the 5′UTR. The PCR products were cleaned using ExoSAP (Affymetrix) and sequenced using BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems) and the manufacturer's protocol. Sequencing reactions were cleaned in Spin-50 mini columns (BioMax, Inc) and resolved on 3100 automated sequencer (Applied Biosystems).

Sequence Analysis and Mutation Discovery

For initial mutation discovery, we sequenced PCR products of *MC1R*, *ASIP,* and *TYRP1* in 4 white, 4 black, and 4 brown dromedaries. Sequences were analyzed for mutations using Sequencher v 5.3 software (Gene Codes Corp.). Effects of single nucleotide changes and indels on protein structure and function were evaluated with Protein Variation Effect Analyzer (PROVEAN) toolkit[6] (Choi et al., 2012; Choi and Chan, 2015). Amino acid sequences of different species were retrieved from NCBI[7] and Ensembl[8]. Comparative analysis of the MC1R protein across species was performed by aligning amino acid sequences in ClustalW (Thompson et al., 1994). We used Transmembrane Protein Topology with a Hidden Markov Model[9] (Moller et al., 2001) to determine MC1R transmembrane domains and evaluate the effect of SNPs; GeneCluster 2.0 (Reich et al., 2004) for comparative analysis of MC1R across species, and ExPASy webtools (Gasteiger et al., 2003) to translate genomic sequence into protein.

[3]http://www.ncbi.nlm.nih.gov

[4]http://genome.ucsc.edu/

[5]http://www.ensembl.org/index.html

[6]http://provean.jcvi.org/index.php

[7]https://www.ncbi.nlm.nih.gov/

[8]http://www.ensembl.org/index.html

[9]http://www.cbs.dtu.dk/services/TMHMM/TMHMM2.0b.guide.php

TABLE 1 | Primers used for PCR and sequencing of *MC1R*, *ASIP,* and *TYRP1.*

Gene symbol and primer ID	Region	Forward 5′ – 3′	Reverse 5′ – 3′	Product size, bp
MC1R 5′ UTR 1	5′ UTR	TCTCACGCCCTTTGAAGTCT	GCTGACGACAAACCCTTCTC	2007
MC1R 5′ UTR 2	5′ UTR	ACACACCTTAACGGGACACC	GCAGGAAAGGGTCTTCACTCT	1101
MC1R 1	Exon 1	TCCTCCTCTGTCTCGTCAGC	GTTGGCTGGCACTGTCTCC	918
MC1R 2	Exon 1	GGCGCTGTCTCTTGTGGAG	GACCAGAAGAGACGCAGGAG	912
MC1R 3	Exon 1	CTCCCTGGCAGGACGATG	AGTCCGAGGTGGGTGGTG	92
*MC1R*_OV*	Exon 1	CATGGTGTCCAGCCTCTGCTCTCT	CACGGCGATAGCACCCAGAGAGCA	n/a
ASIP Prom	Promoter	GGATTTGGGGTCAGTCTGTA	CCCATCCCTTTAGCCTCCTA	2608
ASIP Ex1	Exon 1	GTGTGAGTCAGTGGCAGGAA	AAATTCTGGGTGGGCTAAGG	1130
ASIP Ex1b	Exon 1	ACTTAAGGCAGGCTGGACCT	ATGTGCCCATCCCTTTAGC	1502
*ASIP*ex1seq	Exon 1	AGACCCTGCATTAAGCTGCTC	Sequencing primer	n/a
*ASIP*ex1seqb	Exon 1	GCTTTTCTGATAATGAAATA	Sequencing primer	n/a
ASIP Ex 2	Exon 2	CTTCAGTCTCCCTCCCTTCC	GCCAGGTATTTTTCCCTGAG	828
ASIP Ex 3	Exon 3	TCCAGGGCCTTATTGGACTT	CTGGAAAGGCTCAGTTTGCT	953
ASIP Ex 4	Exon 4	ACTGTAAGAGGGCCAGAGCA	TAAAGTAGGGGGCAGCATTG	511
TYRP1 Ex 1	Exon 1	AGCACTTTGAAGGTGGGTTG	AGTCAGAAGACTGGAGCATCAA	698
TYRP1 Ex 2	Exon 2	AAGAGAGGGAGTGGAAGGGAGA	ATGTGAAATTGCTTGGTCAGTG	650
TYRP1 Ex 3	Exon 3	TGAGTTGGGTTTCATTCCTT	CACTTTCTTTTTCCCCTGGA	629
TYRP1 Ex 4	Exon 4	TGGACATGGTAACTTGGGTTT	GGCCAGCAACCTAACTTTGA	722
TYRP1 Ex 5	Exon 5	GGCCACCAACCATAGGTACA	GACTTCCTGTCTGCCTTTTCA	525
TYRP1 Ex 6	Exon 6	CCTGGGCTGCTGTAGTGAA	CTGGGGGCTCTCAACAAACT	500
TYRP1 Ex 7	Exon 7	GGAATTAGGAAGTGCCCTGA	AACATGCCCCAAATCTTCAC	595

Asterisk denotes overgo primers for screening MC1R from CHORI-246 BAC library.

Large Cohort Genotyping and Association Analysis

Putative causative mutations in *MC1R* and *ASIP* were further analyzed for genotype–phenotype association by Sanger sequencing the regions in 69 dromedaries (29 white, 17 black, 23 brown). Custom TaqMan^{TM} SNP genotyping assays were designed for *MC1R* and *ASIP* mutations according to manufacturer specification (Applied Biosystems) (**Table 2**), and used for genotyping all 188 dromedaries. We used CFX-96 Real Time-PCR machine (Bio-Rad) and corresponding software for PCR amplifications, genotyping and allelic discrimination. The thermal conditions were: priming at 60°C for 1 min, initial denaturation at 95°C for 10 min, 40 cycles of 92°C for 15 s, annealing at primer-specific t°C, extension for 1 min at 60°C, followed by a final extension at 65°C. The 8 μL reactions contained 0.208 μL of TaqMan^{TM} assay, 30 ng template DNA and 4.2 μL of ABI TaqMan Universal Master mix, no UNG (Applied Biosystems).

TABLE 2 | TaqMan assays for genotyping *MC1R* g901C > T and *ASIP* g.174495T > Del.

Primer/probe	5′-3′
MC1R-forward	CTCATCATCTGCAACTCCATCGT
MC1R-reverse	CAGCACCTCTTGGAGTGTCTTC
*MC1R*_VICprobe	ATGCCTTCCGCAGCCA
*MC1R*_FAMprobe	CTATGCCTTCTGCAGCCA
ASIP-forward	CCACTCAGATATCCCAGGATGGA
ASIP-reverse	GCTGTAGGCATTGAGGAAGCA
*ASIP*_VICprobe	CCTCTTCCTAGCTACCC
*ASIP*_FAMprobe	CCTCTTCCAACTACCC

Statistical Analysis

We conducted contingency analysis with JMP program v12 (JMP®, Version 13. SAS Institute Inc., Cary, NC, United States, 1989–2007) to examine the relationship between color phenotypes and genotypes at each variable site. Contingency analysis explores the distribution of a categorical variable Y (color phenotypes) across the level of a second categorical variable X (genotypes). The analysis results include three output files: a mosaic plot, contingency table, and statistical tests (**Supplementary Figure S2**). The mosaic plot is divided into rectangles, so that the vertical length of each rectangle is proportional to the proportions of the observed phenotypes (the Y variable) in each genotype (the X variable). The contingency table is a two-way frequency table with a row for each genotype and a column for each phenotype, and shows their total count, total percent, phenotype percent and genotype percent relative to the total number of observations. The last part of the report shows the results of statistics tests to determine whether or not the phenotypes are independent from genotypes and include R-square and two Chi-square tests, and a probability estimation (Prob > ChiSq) (see **Supplementary Figure S2** for more details).

Chromosome Preparations

Alpaca, dromedary and Bactrian camel chromosome slides were prepared from methanol:acetic acid (3:1)-fixed cell suspensions available in the depository of the Molecular Cytogenetics

laboratory at Texas A&M University. All cell suspensions originated from normal individuals with normal karyotypes.

Fluorescence *in situ* Hybridization (FISH)

We used alpaca CHORI-246 genomic Bacterial Artificial Chromosome (BAC) library[10] to obtain probes for FISH. BAC clones containing *ASIP* and *TYRP1* were previously identified and mapped in the alpaca (Avila et al., 2014a). To obtain BACs for *MC1R*, we screened CHORI-246 filters with *MC1R*-specific radioactively labeled ([^{32}P] dATP/dCTP) overgo primers (**Table 1**) as described by Avila et al. (2014b). The final BACs containing *MC1R* were further verified by PCR with *MC1R* exon primers (**Table 1**). BAC DNA was isolated with Plasmid Mini Kit (Qiagen) according to the manufacturer's protocol. Probe labeling, hybridization and signal detection were conducted according to standard protocols (Raudsepp and Chowdhary, 2008). Because of difficulties to unambiguously identify camelid chromosomes by conventional cytogenetic methods (Avila et al., 2014b), BACs containing the three genes were co-hybridized with a differently labeled reference gene from the alpaca cytogenetic map (Avila et al., 2014a). Composite information about the BACs used for comparative FISH mapping is presented in **Table 3**. Images for at least 10 metaphases for each experiment were captured and analyzed using a Zeiss Axioplan 2 fluorescence microscope, equipped with the Isis Version 5.2 (MetaSystems GmbH) software.

RESULTS

Mutation Discovery and Association Analysis of *MC1R*

The initial sequence analysis of 12 individuals (4 white, 4 black, and 4 brown) identified 7 sequence variants inside and around *MC1R* (**Table 4**). All variants were SNPs and included the previously reported c.901C > T (p.Arg301Cys) missense mutation in the *MC1R* coding region (Almathen et al., 2018) and 6 new non-coding variants: three SNPs in the promoter region, two in 5′UTR and one in 3′UTR. The c.901C > T missense mutation was genotyped in large cohorts by Sanger sequencing (n = 68) and by TaqMan™ genotyping (n = 188) showing that this mutation is significantly associated ($P < 0.0001$) with the white color (**Table 5**), thus confirming the findings of Almathen et al. (2018). To evaluate the possible effect of the p.301R > C mutation on MC1R function, we constructed transmembrane protein topology and showed that the amino acid change affects the last of the 7 transmembrane domains (**Figure 2**). We also aligned the amino acid sequences of the MC1R last transmembrane domain in diverse mammalian and vertebrate species and showed that arginine at this position is highly conserved across species (**Supplementary Figure S1**), suggesting its importance for MC1R normal function.

The initial analysis also indicated that the SNP g.538058G > A in *MC1R* 3′UTR may be associated with color phenotype because genotype GA was present only in black dromedaries (**Table 4**). Large cohort (n = 68) genotyping by sequencing confirmed this and showed that GA genotype was more frequent ($P < 0.0004$; **Table 6**) in black animals. Notably, we did not find dromedaries homozygous for the A-allele (AA) at this site, which is most likely due to the low (0.08) minor-allele (A) frequency, which predicts AA-genotype frequency in small (n = 12) cohort as 0.01 and in large (n = 68) cohort as 0.44 (**Supplementary Table S2**).

Mutation Discovery and Association Analysis of *ASIP*

We identified three sequence variants in the four exons of the dromedary *ASIP* gene: – two previously known in exon 2 (Almathen et al., 2018) and a new frameshift mutation exon 4 (**Table 4**). A single nucleotide deletion in exon 2 (g.174495T_del; c.23T_del) combined with a SNP two base-pairs later (g.174497A > G; c.25A > G; **Table 4**) caused a shift in the reading frame, an insertion of a premature stop at codon 24, and truncated protein (**Figure 3**). All black dromedaries in the discovery cohort were homozygous for the frameshift deletion. Therefore, we genotyped the frameshift mutation in large dromedary cohorts by sequencing (n = 68) and TaqMan™ assay (n = 188). The results showed that the frameshift mutation in exon 2 is significantly associated ($P < 0.0001$) with black coat color (**Table 7**), consistent with the previous findings (Almathen et al., 2018). However, in the large study cohort, one black animal did not have the frameshift deletion and three black dromedaries were heterozygous for it (**Table 7** and **Supplementary Table S1**). In these four animals, we analyzed *ASIP* sequence further and discovered that two dromedaries were heterozygous for another frameshift mutation in exon 4 at g.178388C_del (**Table 4**). The mutation shifted normal stop at codon 133 to codon 254, resulting in 120 amino acids longer polypeptide (**Figure 3**).

[10]http://bacpacresources.org/library.php?id=448

TABLE 3 | Comparative cytogenetic mapping of *ASIP*, *MC1R,* and *TYRP1*.

CHORI-246 BAC	Gene symbol	Camelid chr.	Alpaca	Dromedary	Bactrian	Human chr.
018C13	*ASIP*	19q12	Avila et al., 2014a	This study	This study	20q11.2-q12
125P19*	*EDN3**	19q23	Avila et al., 2014a	This study	This study	20q13.2-q13.3
166N17	*MC1R*	21q15	This study	This study	This study	16q23
128F16*	*MYOC**	21q13	Avila et al., 2014a	This study	This study	1q23-q24
129N17	*TYRP1*	4q21dist-q22	Avila et al., 2014a	This study	This study	9p23
135B22*	*MRPL41**	4q36	Avila et al., 2014a	This study	This study	9q34.3

*Details about alpaca BAC clones, corresponding genes and cytogenetic locations; *denotes reference BACs/genes for chromosome identification.*

TABLE 4 | Sequence polymorphisms in *ASIP*, *MC1R,* and *TYRP1.*

Gene symbol	Location in the gene	Variant	Reference	Effect on protein	Phenotype–genotype			*P-value*
					White	Black	Brown	
MC1R	Promoter (1802 bp from ORF	g.535236C > T	This study	Non-coding	4CC	4CT	4CT	0.0005
MC1R	Promoter (557 bp from ORF)	g.536482G > A	This study	Non-coding	2GG, 2GA	1GG, 2GA, 1AA	2GG, 2GA	0.6208
MC1R	Promoter (420 bp from ORF)	g.536623G > A	This study	Non-coding	4GG	3GG, 1GA	3GG, 1GA	0.4033
MC1R	5′UTR	g.537027G > A	This study	Non-coding	4GG	3GG, 1GA	3GG, 1GA	0.4033
MC1R	5′UTR	g.537028A > T	This study	Non-coding	4AA	3AA, 1AT	3AA, 1AT	0.4033
MC1R	Exon	g.537961C > T; c.901C > T	Almathen et al., 2018	Missense; p.Arg301Cys	3CT, 1TT	4CC	4CC	0.0042
MC1R	3′UTR	g.538058G > A	This study	Non-coding	4GG	2GG, 2GA	4GG	0.0718
ASIP	Exon 2	g.174495T_del; c.23T_del	Almathen et al., 2018	Frameshift; p.24X	4TD	4DD	2TT,2TD	0.0009
ASIP	Exon 2	g.174497A > G; c.25A > G	Almathen et al., 2018	Synonymous	4GA	4AA	2GG, 2GA	0.0009
ASIP	Exon 4	g.178388C_del	This study	Frameshift; p.253X	4CC	2CC, 2CD	3CC, 1CD	0.178
TYRP1	Intron 1 (72 bp before exon 2	g.6544484G > A	This study	Non-coding	4GG	3GG, 1GA	4GG	0.3359
TYRP1	Intron 2 (26 bp after exon 2)	g.6544049T > C	This study	Non-coding	2TT, 2TC	4TT	1TT, 3TC	0.0438
TYRP1	Intron 3 (13 bp before exon 4)	g.65372260_ g.65372261insCA	This study	Non-coding	1insCA/insCA	None	None	0.3359
TYRP1	Intron 4 (90 bp after exon 4	g.6537183T > C	This study	Non-coding	2TT, 2TC	3TT, 1TC	1TT, 3TC	0.3512
TYRP1	3′UTR	g.6529345A > T	This study	Non-coding	3AA, 1AT	3AA, 1AT	3AA, 1AT	1

Sequence variants were discovered by sequencing the three genes in 4 white, 4 black, and 4 brown dromedaries. Sequence positions correspond to dromedary whole genome assembly: GCA_000767585.1 PRJNA234474_Ca_dromedarius_V1.0.; MC1R scaffold ID: NW_011592664.1, ASIP scaffold ID: NW_011591043.1 and TYRP1 scaffold ID: NW_011591511.1; Numbers in columns White, Black, Brown denote the number of animals with the corresponding genotype; ORF, open reading frame; D, deletion; P-value for genotype–phenotype association was determined by contingency analysis in JMP using Chi-square (see ***Supplementary Figure S2****).*

TABLE 5 | Genotype frequencies of *MC1R* c.901C > T missense mutation in a large study cohort (*n* = 188).

Genotypes	Frequency (count) in dromedary color groups		
	White (*n* = 53)	Black (*n* = 38)	Brown (*n* = 97)
CC	0.037 (7)	0.20 (38)	0.41 (78)
CT	0.13 (25)	0	0.085 (16)
TT	0.11 (21)	0	0.016 (3)

The mutation is significantly (P < 0.0001) associated with white coat color.

TABLE 6 | Genotype frequencies of *MC1R* 3′UTR variant g.538058G > A in a large study cohort (*n* = 68).

Genotypes	Frequency (count) in dromedary color groups		
	White (*n* = 22)	Black (*n* = 15)	Brown (*n* = 31)
GG	0.29 (20)	0.13 (9)	0.46 (31)
GA	0.029 (2)	0.088 (6)	0
AA	0	0	0

The SNP is significantly (P < 0.0004) associated with black coat color.

However, the other two animals did not have this deletion and, overall, we were not able to associate exon 4 mutation with black color in our study cohort.

Mutation Discovery in *TYRP1*

Sequence analysis of the 7 exons and exon–intron boundaries of the *TYRP1* gene in the discovery cohort of 12 dromedaries, identified 5 sequence variants: 4 SNPs and one insertion (**Table 4**). However, all variants were in non-coding regions (introns and 3′UTR) and not associated with dromedary color phenotypes. Therefore, we did not conduct any large cohort genotyping for *TYRP1*.

Comparative FISH Mapping

We mapped *TYRP1, ASIP,* and *MC1R* by FISH to metaphase chromosomes in the alpaca, dromedary and Bactrian camel (**Figure 4**). For unambiguous chromosome identification, we used chromosome-specific reference markers from the alpaca cytogenetic map (Avila et al., 2014a). The *TYRP1* and *ASIP* genes were previously FISH mapped to alpaca chromosomes 4 and 19, respectively (Avila et al., 2014a). Here we mapped *TYRP1* to chr4 and *ASIP* to chr19 in both camel species (**Figures 4A,B**). The results are in agreement with karyotype conservation across Old and New World camelids (Bunch et al., 1985; Bianchi et al., 1986) and consistent with human-dromedary Zoo-FISH data on conserved synteny segments between these species (Balmus et al., 2007). The *MC1R* gene has not been chromosomally assigned in any camelid genome. Here we mapped *MC1R* to the very terminal region in chr21q in the alpaca, dromedary and Bactrian camel (**Figure 4C**) and revealed a hitherto unknown conserved synteny block between camelid chr21 and HSA16q.

DISCUSSION

Here we validated the recently published mutations for white and black/dark brown coat color in dromedaries (Almathen et al., 2018) using independent dromedary populations of US and Saudi Arabian origin. In addition, we designed for both mutations TaqMan^{TM} assays and confirmed their accuracy and efficiency for large cohort genotyping, suggesting that the high throughput, faster and cheaper TaqMan^{TM} assay should be the method of choice for any further genotyping of validated color-associated variants in large numbers of additional animals.

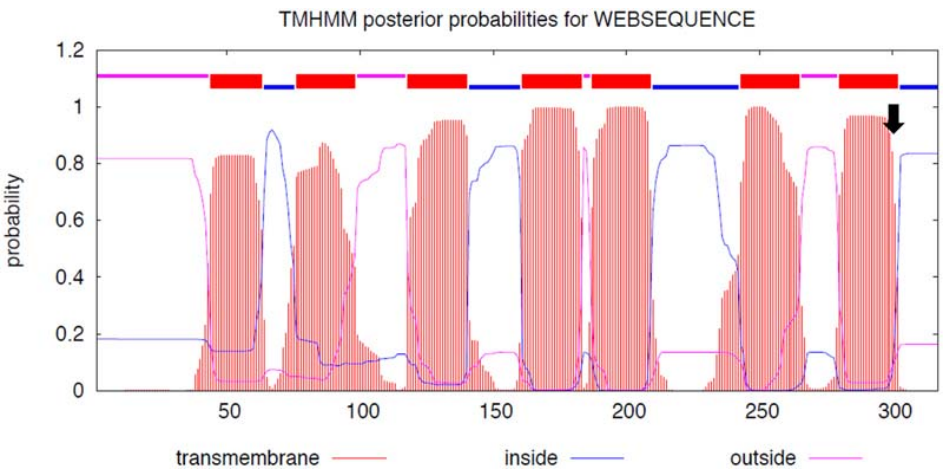


FIGURE 2 | MC1R protein functional domains. The seven MC1R transmembrane domains and the position of p.301R > C mutation (arrow); *y*-axis: the probability of the amino acid sequences to be cytoplasmic (blue), extracellular (magenta), or part of the transmembrane helix (orange); *x*-axis: amino acid sequence. We used Transmembrane Protein Topology with a Hidden Markov Model (Moller et al., 2001; http://www.cbs.dtu.dk/services/TMHMM/TMHMM2.0b.guide.php).

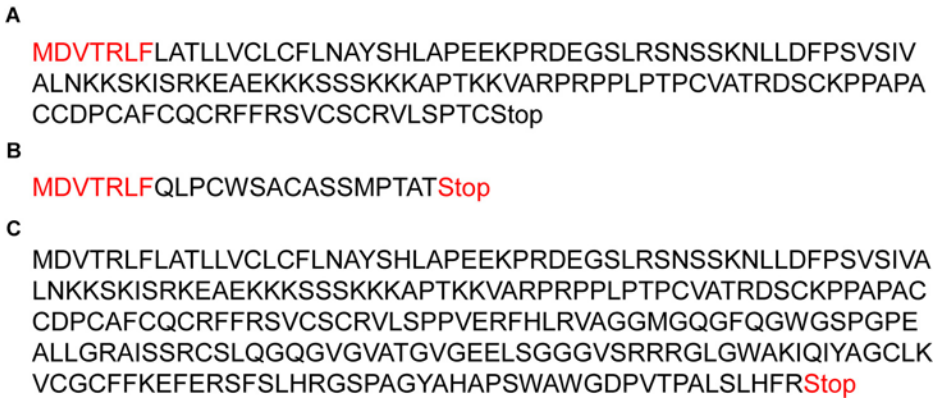


FIGURE 3 | The effect of frameshift mutations on ASIP polypeptide. **(A)** Normal ASIP polypeptide with 133 amino acids and stop at codon 134; **(B)** Truncated ASIP protein with 24 amino acids and stop at codon 25 due to frameshift mutation in exon 2; **(C)** Abnormally long polypeptide with 253 amino acids due to a frameshift mutation in exon 4. Amino acids in red font in **(A,B)** are before frameshift and, thus shared between the normal and truncated ASIP.

Overall, our results were consistent with the recently published data, but also refined and expanded it. Besides confirming the previously reported single variant in *MC1R* and two variants in *ASIP* (Almathen et al., 2018), we found six additional SNPs in *MC1R* non-coding regions, a new deletion in *ASIP* exon 4, and 5 novel variants in *TYRP1*.

The likely causative mutation for the white color in *MC1R* c.901C > T in dromedaries was, at the first sight intriguing because, according to published references (Rieder et al., 2001; Almathen et al., 2018), a mutation at the same position (c.901C > T) is responsible for recessive chestnut coat color in horses. This poses a question why the same missense mutation results in a depigmentation phenotype in the dromedary, but a pheomelanic phenotype in horses. However, closer inspection of the original publication for the horse chestnut mutation (Marklund et al., 1996; reviewed by Andersson, 2003) reveals that the horse chestnut and dromedary white mutations are different. The horse chestnut is due to p.Ser83Phe, which affects MC1R second transmembrane domain (Marklund et al., 1996), while the dromedary mutation p.Arg301Cys is in the last (seventh) transmembrane domain (**Figure 2**).

The dromedary mutation, however, shares functional and phenotypic similarity to recently reported *MC1R* sequence variants in Australian cattle dogs and Alaskan and Siberian huskies (Durig et al., 2018). Cream color in Australian cattle dogs is associated with a combination of c.916C > T (p.Arg306Ter) and a promoter variant affecting MITF binding site. White huskies, on the other hand, are homozygous for a deletion c.816-delCT. Even though causative sequence variants are different in Australian cattle dogs, huskies and dromedaries, they share essential similarities: all occur in the last transmembrane domain, negatively affect MC1R function and result in depigmentation phenotypes. The mutations in cream-colored Australian cattle dogs cause downregulation of *MC1R* transcription, while MC1R in white huskies has lost the last transmembrane domain and the cytoplasmic C-terminal tail (Durig et al., 2018). Though no functional data are available for white dromedaries, we theorize based on the predicted effect of p.Arg301Cys on the last transmembrane domain (**Figure 2**) and the resulting white phenotype, that the mutation is loss-of-function. Functional importance of this portion of the MC1R protein is illustrated by highly conserved sequence of 17 amino acids (p.296–312) across diverse mammalian species (**Supplementary Figure S1**). Notably, the dromedary differs from other mammals at p.301 because a white and not a wild-type animal was used for the reference sequence[11].

In contrast to huskies where white color is a recessive trait (Durig et al., 2018), the dromedary white mutation is dominant because heterozygosity for the T-allele at c901C > T is sufficient for the white phenotype (**Table 4** and **Supplementary Table S1**). Therefore, we suggest that the *MC1R* mutation in white dromedaries has dominant negative effect, i.e., it alters the function of the wild-type C-allele and has dominant or semi-dominant phenotype. Similar dominant negative effect on wild-type MC1R receptor cell surface expression or wild-type MC1R cAMP signaling has been described for several *MC1R* sequence variants in humans (Beaumont et al., 2007).

Another observation about dromedary *MC1R*, as also noted by Almathen et al. (2018), is the low level of sequence variation (just c901C > T) in the coding region, contrasting the 21 sequence

TABLE 7 | Genotype frequencies of *ASIP* exon 2 g.174495T_del (D) nonsense mutation in a large study cohort (*n* = 188).

Genotypes	Frequency (count) in dromedary color groups		
	White (*n* = 53)	Black (*n* = 38)	Brown (*n* = 97)
TT	0.12 (23)	0.005 (1)	0.20 (37)
TD	0.13 (25)	0.016 (3)	0.18 (33)
DD	0.03 (5)	0.18 (34)	0.14 (27)

The mutation is significantly (P < 0.0001) associated with black coat color.

[11] https://www.ncbi.nlm.nih.gov/genome/10740?genome_assembly_id=212034

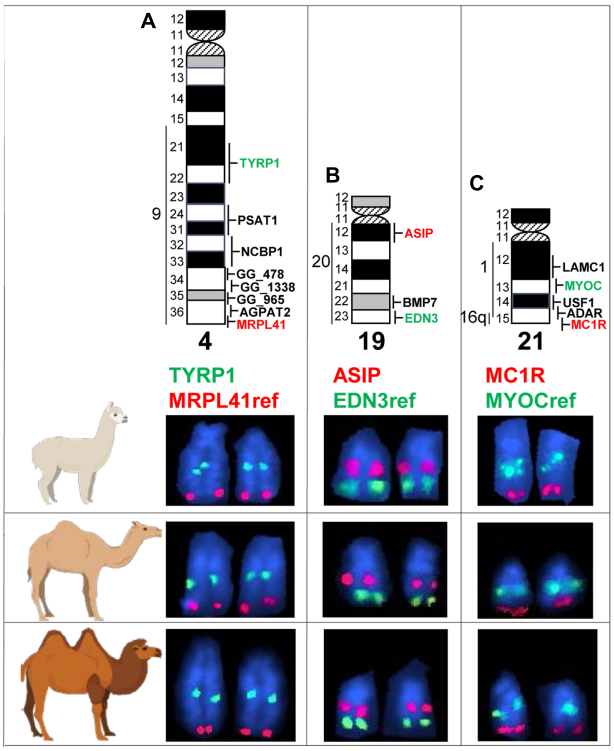


FIGURE 4 | Comparative FISH mapping. Comparative mapping of *TYRP1* **(A)**, *ASIP* **(B)**, and *MC1R* **(C)** in alpaca, dromedary, and Bactrian camel chr4, chr19, and chr21, respectively. Chromosome ideograms with all mapped markers (Avila et al., 2014a) are shown at the top. Vertical lines with numbers to the left of chromosome ideograms indicate homology segments to human chromosomes. Ref, reference gene for chromosome identification. Green and red font colors for coat color genes and reference genes correspond to the green and red FISH signals in partial microscope images below ideograms.

variants found in the alpaca *MC1R* (Feeley and Munyard, 2009). However, a recent study of *MC1R* sequence variants across all four South American camelids (vicugna, guanaco, llama, and alpaca) suggests that variation in alpacas is the result of human selection for a variety of fiber colors, whereas in wild South American camelids (guanacos and free living vicugnas), there is a selection against non-synonymous substitutions in *MC1R* (Marin et al., 2018). Likewise, there is low sequence variation of *MC1R* in wild pigs, but many more variants in domestic pig breeds as a result of human selection (Andersson, 2003). Thus, we suggest that low sequence variation of *MC1R* in dromedaries is because human selection for coat color in this species is a more recent event in course of domestication.

On the other hand, sequence variants are present immediately outside the dromedary *MC1R* coding region, in 5′- and 3′-UTRs and in the promoter (**Table 4**). Whether any of these have regulatory roles in shaping pigmentation phenotypes, is a subject of future studies. This also applies to the 3′UTR variant g.538058G > A (**Table 4**), which showed association ($P < 0.0004$) with black coat color (**Table 6**). Though, it is also possible that the statistical significance may be influenced by relatedness between black animals.

The causative mutation for black coat color, as reported earlier (Almathen et al., 2018) and confirmed in this study (**Table 7**), is a frameshift deletion in *ASIP* exon 2, resulting in premature stop codon and truncated protein (**Figure 3**). Like in previous study (Almathen et al., 2018), we also observed a synonymous SNP 2 bp after the frameshift deletion (**Table 4**), but did not conduct association analysis because it was irrelevant for the premature stop codon. Similar, though not identical, loss-of-function mutations in *ASIP* underlie recessive black color in several domestic and wild species. For example, in alpacas (Feeley et al., 2011), sheep (Norris and Whan, 2008; Royo et al., 2008), Iranian Markhoz goats (Nazari-Ghadikolaei et al., 2018), donkeys (Abitbol et al., 2015), horses (Rieder et al., 2001), dogs (Kerns et al., 2004), cats (Eizirik et al., 2003), and impala antelope (Miller et al., 2016). Like in these species, we are confident that the black color in the dromedary is a recessive trait because the majority (34/38) of black dromedaries in this study were homozygous for the deletion (**Table 7**). However, 4 black animals in our study cohort did not follow this pattern (**Table 7**). Two of these carried another frameshift deletion in *ASIP* exon 4, resulting in abnormally long and likely non-functional ASIP protein (**Figure 3**). We suggest that the second frameshift deletion may be causative for black color in the absence of the first deletion, though it was not possible to conduct association analysis with just 2 individuals. Of the remaining two black dromedaries, one was heterozygous for the exon 2 deletion and the other had no mutations in *ASIP*. This is similar to observations in alpacas where homozygous recessive loss-of-function mutations in *ASIP* explain the majority but not all cases of the black phenotype (Feeley et al., 2011). Thus, like in alpacas, black coat color in dromedaries may be influenced by additional regulatory mutations and MC1R interactions with ASIP and α-melanocyte stimulating hormone (α-MSH). Nevertheless, at this point we did not conduct multi-locus testing because the majority of novel variants were non-coding, and because 12 animals in the discovery cohort would not give enough statistical power for these analyses. Besides, one should also consider possible errors in phenotyping.

We investigated the *TYRP1* gene as a possible contributor to various shades of brown coat color in the dromedary. The gene encodes for an important enzyme for the synthesis of eumelanin (del Marmol and Beermann, 1996) and *TYRP1* mutations are associated with brown or chocolate coat color on black genetic background in many mammals and other vertebrates (see Li et al., 2018). However, all *TYRP1* variants found in this study, were in non-coding regions (**Table 4**) and we did not detect the two SNPs in dromedary *TYRP1* exon1 as reported by a prior study (see Almathen et al., 2018). Nevertheless, both the non-coding SNPs and the exon 1 SNPs were not associated with any color phenotypes. Likewise, no candidate coat color mutations have been detected in alpaca *TYRP1* (Cransberg and Munyard, 2011). Despite these findings, *TYRP1* remains an important candidate gene for color phenotypes in camelids and should be included in future studies.

Finally, we comparatively FISH mapped the three coat color genes in three camelid species – the alpaca, the dromedary, and

the Bactrian camel. In agreement with the known conservation of camelid karyotypes (Taylor et al., 1968; Bunch et al., 1985) and prior mapping of *TYRP1* and *ASIP* in alpacas (Avila et al., 2014a), the genes mapped to the same cytogenetic location in the same chromosomes in all species: *TYRP1* to chr4q21-q22, *ASIP* to chr19q12, and *MC1R* to chr21qter (**Figure 4**). While the locations of *TYRP1* and *ASIP* were in good agreement with human-dromedary Zoo-FISH (Balmus et al., 2007), mapping *MC1R* to chr21 came as a surprise. This is because camelid chr21 shares known conserved synteny with part of HSA1q only (Balmus et al., 2007; Avila et al., 2014a). Since human *MC1R* is located very terminal in the long arm of chr16 (HSA16q24.3; 89.9 Mb)[12], we anticipated mapping *MC1R* to camelid chr9, which is homologous to HSA16q (Balmus et al., 2007; Avila et al., 2014a). Furthermore, camelid chr9 shares also homology with HSA19q, and HSA16q/HSA19q correspond to an ancestral eutherian synteny combination, which has been conserved in many eutherian karyotypes (Chowdhary et al., 1998; Ferguson-Smith and Trifonov, 2007). Our findings indicate that this ancestral synteny combination has undergone rearrangements during camelid karyotype evolution, so that a segment homologous to HSA16q containing *MC1R* has become a part of camelid chr21 and shares synteny with sequences corresponding to HSA1q. Inspection of the current dromedary genome assembly PRJNA234474_Ca_dromedarius_V1.0[13] scaffolds confirmed FISH results for *MC1R* and showed that sequences corresponding to HSA1q: 145–147 Mb and HSA16q: 85–90 Mb are together in dromedary scaffold479 sequence NW_011591415.1[14]. Therefore, cytogenetic mapping of *MC1R* in camelids revealed a novel human-camelid synteny segment, confirmed sequence assembly of scaffold479, and anchored alpaca, dromedary and Bactrian camel scaffolds containing *TYRP1*, *ASIP*, and *MC1R* to chromosomes.

ETHICS STATEMENT

Procurement of peripheral blood was performed according to the United States Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training. All procedures were approved by Institutional Animal Care and Use Committee as AUP#2011-96, AUP#2018-0342CA and CRRC#09-47 at Texas A&M University.

[12] https://genome.ucsc.edu/

[13] https://www.ncbi.nlm.nih.gov/assembly/GCF_000767585.1/

[14] https://www.ncbi.nlm.nih.gov/nuccore/NW_011591415.1?report=graph

AUTHOR CONTRIBUTIONS

TR and FA initiated and designed the study. FA, CC, RJ, AH, MM, GG, and FPdL conducted the experimental work and data analysis. TR, FA, and CC wrote the manuscript with input from all authors.

FUNDING

This study was supported by grants from Alpaca Research Foundation (ARF) 2009–2011 and Morris Animal Foundation (MAF) D09LA-004, D14LA-005. The authors highly appreciate donations to ARF and MAF by Leslie Herzog of Herzog Alpacas. FA was supported by Qassim University, Saudi Arabia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00340/full#supplementary-material

FIGURE S1 | *MC1R* comparative. Comparative alignment of *MC1R* transmembrane domain 7 amino acid sequences in diverse mammalian and vertebrate species. The p.301 position is indicated by a vertical arrow and highlighted; horizontal arrow shows the p.301R > C mutation in the dromedary. Note that Arginine is highly conserved across species, except in humans who have Histidine, which is another positively charged amino acid.

FIGURE S2 | Contingency analysis output files in JMP program for the sequence variants analyzed in this study. Each output file comprises a Mosaic Plot, a Contingency Table, and Statistics Tests; N, the total number of observations; DF, records the degrees of freedom associated with the test; -LogLike, negative log-likelihood; Rsquare (U), shows portion of the total uncertainty attributed to the model fit. ChiSquare, two chi-square statistical tests; Prob > ChiSq, lists the probability of obtaining, by chance alone, a Chi-square value greater than the one computed if no relationship exists between phenotype and genotype.

TABLE S1 | Composite information about the dromedaries used in this study (color, ID, country of origin, sex, photo, *MC1R* g901C > T genotype; *ASIP* exon 2 g.174495T > Del nonsense mutation, and *MC1R* 3′UTRG > A genotype); * denote the animals for which genotypes were determined both by Sanger sequencing and TaqMan™ assay.

TABLE S2 | Expected and observed allele and genotype frequencies for all variants found in this study.

REFERENCES

Abitbol, M., Legrand, R., and Tiret, L. (2015). A missense mutation in the agouti signaling protein gene (*ASIP*) is associated with the no light points coat phenotype in donkeys. *Genet. Sel. Evol.* 47:28. doi: 10.1186/s12711-015-0112-x

Almathen, F., Elbir, H., Bahbahani, H., Mwacharo, J., and Hanotte, O. (2018). Polymorphisms in *MC1R* and *ASIP* genes are associated with coat color variation in the arabian camel. *J. Hered.* 109, 700–706. doi: 10.1093/jhered/esy024

Andersson, L. (2001). Genetic dissection of phenotypic diversity in farm animals. *Nat. Rev. Genet.* 2, 130–138. doi: 10.1038/35052563

Andersson, L. (2003). Melanocortin receptor variants with phenotypic effects in horse, pig, and chicken. *Ann. N. Y. Acad. Sci.* 994, 313–318. doi: 10.1111/j.1749-6632.2003.tb03195.x

Avila, F., Baily, M. P., Perelman, P., Das, P. J., Pontius, J., Chowdhary, R., et al. (2014a). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*Lama pacos*). *Cytogenet. Genome Res.* 144, 196–207. doi: 10.1159/000370329

Avila, F., Das, P. J., Kutzler, M., Owens, E., Perelman, P., Rubes, J., et al. (2014b). Development and application of camelid molecular cytogenetic tools. *J. Hered.* 105, 858–869. doi: 10.1093/jhered/ess067

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C., Alkalaeva, E. S., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative cetartiodactyla ancestral karyotype. *Chromosome Res.* 15, 499–515. doi: 10.1007/s10577-007-1154-x

Beaumont, K. A., Shekar, S. N., Newton, R. A., James, M. R., Stow, J. L., Duffy, D. L., et al. (2007). Receptor function, dominant negative activity and phenotype correlations for *MC1R* variant alleles. *Hum. Mol. Genet.* 16, 2249–2260. doi: 10.1093/hmg/ddm177

Bellone, R. R. (2010). Pleiotropic effects of pigmentation genes in horses. *Anim. Genet.* 41(Suppl. 2), 100–110. doi: 10.1111/j.1365-2052.2010.02116.x

Bianchi, N. O., Larramendy, M. L., Bianchi, M. S., and Cortes, L. (1986). Karyological conservation in south american camelids. *Experientia* 42, 622–624. doi: 10.1007/BF01955563

Bunch, T. D., Foote, W. C., and Maciulis, A. (1985). Chromosome banding pattern homologies and NORs for the bactrian camel, guanaco, and llama. *J. Hered.* 76, 115–118. doi: 10.1093/oxfordjournals.jhered.a110034

Chandramohan, B., Renieri, C., La Manna, V., and La Terza, A. (2013). The alpaca agouti gene: genomic locus, transcripts and causative mutations of eumelanic and pheomelanic coat color. *Gene* 521, 303–310. doi: 10.1016/j.gene.2013.03.060

Chandramohan, B., Renieri, C., La Manna, V., and La Terza, A. (2015). The alpaca melanocortin 1 receptor: gene mutations, transcripts, and relative levels of expression in ventral skin biopsies. *ScientificWorldJournal* 2015:265751. doi: 10.1155/2015/265751

Choi, Y., and Chan, A. P. (2015). Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31, 2745–2747. doi: 10.1093/bioinformatics/btv195

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7:e46688. doi: 10.1371/journal.pone.0046688

Chowdhary, B. P., Raudsepp, T., Fronicke, L., and Scherthan, H. (1998). Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome Res.* 8, 577–589. doi: 10.1101/gr.8.6.577

Cieslak, M., Reissmann, M., Hofreiter, M., and Ludwig, A. (2011). Colours of domestication. *Biol. Rev. Camb. Philos. Soc.* 86, 885–899. doi: 10.1111/j.1469-185X.2011.00177.x

Cransberg, R., and Munyard, K. A. (2011). Polymorphisms detected in the tyrosinase and *MATP* (*SLC45A2*) genes did not explain coat colour dilution in a sample of alpaca (*Viguna pacos*). *Small Rumin. Res.* 95, 92–96. doi: 10.1016/j.smallrumres.2010.10.004

del Marmol, V., and Beermann, F. (1996). Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* 381, 165–168. doi: 10.1016/0014-5793(96)00109-3

Durig, N., Letko, A., Lepori, V., Hadji Rasouliha, S., Loechel, R., Kehl, A., et al. (2018). Two *MC1R* loss-of-function alleles in cream-coloured Australian cattle dogs and white huskies. *Anim. Genet.* 49, 284–290. doi: 10.1111/age.12660

Eizirik, E., Yuhki, N., Johnson, W. E., Menotti-Raymond, M., Hannah, S. S., and O'Brien, S. J. (2003). Molecular genetics and evolution of melanism in the cat family. *Curr. Biol.* 13, 448–453. doi: 10.1016/S0960-9822(03)00128-3

Feeley, N. L., Bottomley, S., and Munyard, K. A. (2011). Three novel mutations in *ASIP* associated with black fibre in alpacas (*Vicugna pacos*). *J. Agric. Sci.* 149, 529–538. doi: 10.1017/S0021859610001231

Feeley, N. L., and Munyard, K. A. (2009). Characterisation of the melanocortin-1 receptor gene in alpaca and identification of possible markers associated with phenotypic variations in colour. *Anim. Prod. Sci.* 49, 675–681. doi: 10.1071/AN09005

Ferguson-Smith, M. A., and Trifonov, V. (2007). Mammalian karyotype evolution. *Nat. Rev. Genet.* 8, 950–962. doi: 10.1038/nrg2199

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003). Expasy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788. doi: 10.1093/nar/gkg563

Guridi, M., Soret, B., Alfonso, L., and Arana, A. (2011). Single nucleotide polymorphisms in the melanocortin 1 receptor gene are linked with lightness of fibre colour in peruvian alpaca (*Vicugna pacos*). *Anim. Genet.* 42, 679–682. doi: 10.1111/j.1365-2052.2011.02205.x

Haldane, J. B. S. (1927). The comparative genetics of colour in rodents and carnivora. *Biol. Rev.* 2, 199–210. doi: 10.1111/j.1469-185X.1927.tb00877.x

Holl, H., Isaza, R., Mohamoud, Y., Ahmed, A., Almathen, F., Youcef, C., et al. (2017). A frameshift mutation in *KIT* is associated with white spotting in the Arabian camel. *Genes* 8:102. doi: 10.3390/genes8030102

Kerns, J. A., Newton, J., Berryere, T. G., Rubin, E. M., Cheng, J. F., Schmutz, S. M., et al. (2004). Characterization of the dog agouti gene and a non agouti mutation in German shepherd dogs. *Mamm. Genome* 15, 798–808. doi: 10.1007/s00335-004-2377-1

Li, J., Bed'hom, B., Marthey, S., Valade, M., Dureux, A., Moroldo, M., et al. (2018). A missense mutation in *TYRP1* causes the chocolate plumage color in chicken and alters melanosome structure. *Pigment Cell Melanoma Res.* doi: 10.1111/pcmr.12753 [Epub ahead of print].

Marin, J. C., Rivera, R., Varas, V., Cortes, J., Agapito, A., Chero, A., et al. (2018). Genetic variation in coat colour genes *MC1R* and *ASIP* provides insights into domestication and management of south american camelids. *Front. Genet.* 9:487. doi: 10.3389/fgene.2018.00487

Marklund, L., Moller, M. J., Sandberg, K., and Andersson, L. (1996). A missense mutation in the gene for melanocyte-stimulating hormone receptor (*MC1R*) is associated with the chestnut coat color in horses. *Mamm. Genome* 7, 895–899. doi: 10.1007/s003359900264

Miller, S. M., Guthrie, A. J., and Harper, C. K. (2016). Single base-pair deletion in *ASIP* exon 3 associated with recessive black phenotype in impala (*Aepyceros melampus*). *Anim. Genet.* 47, 511–512. doi: 10.1111/age.12430

Moller, S., Croning, M. D., and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646–653. doi: 10.1093/bioinformatics/17.7.646

Morante, R., Goyache, F., Burgos, A., Cervantes, I., Pérez-Cabal, M. A., and Gutiérrez, J. P. (2009). Genetic improvement for alpaca fibre production in the peruvian altiplano: the pacomarca experience. *Anim. Genet. Resourc. Inform.* 45, 37–43. doi: 10.1017/S1014233909990307

Nazari-Ghadikolaei, A., Mehrabani-Yeganeh, H., Miarei-Aashtiani, S. R., Staiger, E. A., Rashidi, A., and Huson, H. J. (2018). Genome-wide association studies identify candidate genes for coat color and mohair traits in the Iranian Markhoz goat. *Front. Genet.* 9:105. doi: 10.3389/fgene.2018.00105

Norris, B. J., and Whan, V. A. (2008). A gene duplication affecting expression of the ovine *ASIP* gene is responsible for white and black sheep. *Genome Res.* 18, 1282–1293. doi: 10.1101/gr.072090.107

Pielberg, G. (2004). *Molecular Coat Color Genetics: Department of Animal Breeding and Genetics*. Ph.D. dissertation, Swedish University of Agricultural Sciences, Uppsala, 34.

Raudsepp, T., and Chowdhary, B. P. (2008). FISH for mapping single copy genes. *Methods Mol. Biol.* 422, 31–49. doi: 10.1007/978-1-59745-581-7_3

Rees, J. L. (2003). Genetics of hair and skin color. *Annu. Rev. Genet.* 37, 67–90. doi: 10.1146/annurev.genet.37.110801.143233

Reich, M., Ohm, K., Angelo, M., Tamayo, P., and Mesirov, J. P. (2004). Genecluster 2.0: an advanced toolset for bioarray analysis. *Bioinformatics* 20, 1797–1798. doi: 10.1093/bioinformatics/bth138

Reissmann, M., and Ludwig, A. (2013). Pleiotropic effects of coat colour-associated mutations in humans, mice and other mammals. *Semin. Cell Dev. Biol.* 24, 576–586. doi: 10.1016/j.semcdb.2013.03.014

Rieder, S., Taourit, S., Mariat, D., Langlois, B., and Guerin, G. (2001). Mutations in the agouti (*ASIP*), the extension (*MC1R*), and the brown (*TYRP1*) loci and their association to coat color phenotypes in horses (*Equus caballus*). *Mamm. Genome* 12, 450–455. doi: 10.1007/s003350020017

Royo, L. J., Alvarez, I., Arranz, J. J., Fernandez, I., Rodriguez, A., Perez-Pardal, L., et al. (2008). Differences in the expression of the *ASIP* gene are involved in the recessive black coat colour pattern in sheep: evidence from the rare Xalda sheep breed. *Anim. Genet.* 39, 290–293. doi: 10.1111/j.1365-2052.2008.01712.x

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Schmutz, S. M., and Berryere, T. G. (2007). Genes affecting coat colour and pattern in domestic dogs: a review. *Anim. Genet.* 38, 539–549. doi: 10.1111/j.1365-2052.2007.01664.x

Sturm, R. A., and Duffy, D. L. (2012). Human pigmentation genes under environmental selection. *Genome Biol.* 13:248. doi: 10.1186/gb-2012-13-9-248

Suzuki, H. (2013). Evolutionary and phylogeographic views on *MC1R* and *ASIP* variation in mammals. *Genes Genet. Syst.* 88, 155–164. doi: 10.1266/ggs.88.155

Taylor, K. M., Hungerford, D. A., Snyder, R. L., and Ulmer, F. A. Jr. (1968). Uniformity of kryotypes in the camelidae. *Cytogenetics* 7, 8–15. doi: 10.1159/000129967

Thompson, J. D., Higgins, D. G., Gibson, T. J., and Clustal, W. (1994). Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, e115. doi: 10.1093/nar/gks596

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor and reviewer PO-tW declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

Copyright © 2019 Alshanbari, Castaneda, Juras, Hillhouse, Mendoza, Gutiérrez, Ponce de León and Raudsepp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fgene.2019.00361

Evaluation of SNP Genotyping in Alpacas Using the Bovine HD Genotyping Beadchip

*Manuel More[1], Gustavo Gutiérrez[1], Max Rothschild[2], Francesca Bertolini[3] and F. Abel Ponce de León[4]**

[1] Facultad de Zootecnia, Universidad Nacional Agraria La Molina, Lima, Peru, [2] Department of Animal Science, Iowa State University, Ames, IA, United States, [3] National Institute of Aquatic Resources, DTU-Aqua, Technical University of Denmark, Lyngby, Denmark, [4] Department of Animal Science, University of Minnesota, Minneapolis, MN, United States

Alpacas are one of four South American Camelid species living in the highlands of the Andes. Production of alpaca fiber contributes to the economy of the region and the livelihood of many rural families. Fiber quantity and quality are important and in need of a modern breeding program based on genomic selection to accelerate genetic gain. To achieve this is necessary to discover enough molecular markers, single nucleotide polymorphisms (SNPs) in particular, to provide genome coverage and facilitate genome wide association studies to fiber production characteristics. The aim of this study was to discover alpaca SNPs by genotyping forty alpaca DNA samples using the BovineHD Genotyping Beadchip. Data analysis was performed with GenomeStudio (Illumina) software. Because different filters and thresholds are reported in the literature we investigated the effects of no-call threshold (≥0.05, ≥0.15, and ≥0.25) and call frequency (≥0.9 and =1.0) in identifying positive SNPs. Average GC Scores, calculated as the average of the 10% and 50% GenCall scores for each SNP (≥0.70) and the GenTrain score ≥ 0.25 parameters were applied to all comparisons. SNPs with minor allele frequency (MAF) ≥ 0.05 or ≥ 0.01 were retained. Since detection of SNPs is based on the stable binding of oligonucleotide probes to the target DNA immediately adjacent to the variant nucleotide, all positive SNP flanking sequences showing perfect alignments between the bovine and alpaca genomes for the first 21 or 26 nucleotides flanking the variant nucleotide at either side were selected. Only SNPs localized in one scaffold were assumed unique. Unique SNPs identified in both reference genomes were kept and mapped on the Vicugna_pacos 2.0.2 genome. The effects of the no-call threshold ≥ 0.25, call frequency = 1 and average GC ≥ 0.7 were meaningful and identified 6756 SNPs of which 400 were unique and polymorphic (MAF ≥ 0.01). Assignment to alpaca chromosomes was possible for 292 SNPs. Likewise, 209 SNPs were localized in 202 alpaca gene loci and 29 of these share the same loci with the dromedary. Interestingly, 69 of 400 alpaca SNPs have 100% similarity with dromedary.

Keywords: alpaca, bovine, SNP, genotyping, polymorphic

OPEN ACCESS

Edited by:
Pamela Burger, University of Veterinary Medicine, Austria

Reviewed by:
Laura B. Scheinfeldt, University of Pennsylvania, United States
Felipe Avila, University of California, Davis, United States

***Correspondence:**
F. Abel Ponce de León
apl@umn.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *28 October 2018*
Accepted: *04 April 2019*
Published: *24 April 2019*

Citation:
More M, Gutiérrez G, Rothschild M, Bertolini F and Ponce de León FA (2019) Evaluation of SNP Genotyping in Alpacas Using the Bovine HD Genotyping Beadchip. Front. Genet. 10:361. doi: 10.3389/fgene.2019.00361

INTRODUCTION

Alpacas are an important animal resource living in the highland areas of the Andes. They provide fiber, skins, meat and manure for agricultural production and, along with llamas, are a cornerstone of cultural heritage. Peru hosts about 85% of the worldwide alpaca population of which 80% belong to the Huacaya type, 12% to the Suri type and 8% are intermediate Ministerio de Agricultura y Riego [MINAGRI] (2017). Alpacas are kept mainly for fiber production and meat is a secondary product. Production of alpaca fiber contributes to the regional economy and is in high demand by the textile industry. In 2015 fiber production reached 4,478t at national level, of which 90% was for export market and 10% for the Peruvian market. Individual alpaca breeding program initiatives by private companies, NGOs and farmer cooperatives aimed to improve fiber quality by reducing fiber diameter. Much could be gained with the application of genomic selection to accelerate genetic gain. However, there is still limited information about the alpaca genome organization and a paucity in developing molecular markers necessary for the application of modern animal selection programs.

Several advances in the understanding of the organization of the alpaca genome have occurred in the last decade. The alpaca genome has been sequenced by two separate research groups at a depth of ∼22X (Warren et al., 2013) and 72.5X (Wu et al., 2014). Their corresponding genome assemblies are publicly available. Similarly, chromosomal identification of syntenic regions between human, bovine and camelid by Zoo-FISH have allowed the preliminary assignment of alpaca genome scaffolds to specific alpaca chromosomes (Balmus et al., 2007). Avila et al. (2014) extended the latter, by developing the first cytogenetic map containing 230 chromosomally localized molecular markers and genes. However, there is still a limited number of available molecular markers (Pérez-Cabal et al., 2010; Paredes et al., 2014) and subsequently a very limited number of association studies of genetic markers to production traits in alpacas have been performed (Guridi et al., 2011; Paredes et al., 2014; Chandramohan et al., 2015). Therefore the identification of additional single nucleotide polymorphisms (SNPs) is necessary to improve the SNP coverage across the genome (Munyard et al., 2009), to increase the possibility of identifying linkage disequilibrium between markers and therefore to perform genome-wide association analyses with production traits (Hayes and Goddard, 2010; Dekkers, 2012).

The lack of SNP microarrays for non-model organisms has led to test commercially available SNP microarrays of closely related species to discover common SNPs. Slate et al. (2009) have reviewed alternatives to cross-species application of commercial SNP chips for SNP discovery. Most are labor intensive, high cost, and yield low numbers of SNP in comparison to genotype-by-sequencing (GBS) methods that yield abundant species-specific SNPs at low cost (Miller et al., 2012). However, GBS is prone to higher calling rate errors than genotyping with SNP chips because it relies on pooling random sequence information from several individuals and loci increasing the probability of low coverage for some individual/locus combinations. SNP chips, on the other hand, have the advantage that each locus is present multiple times in the chip and genotypes are called by averaging over all of the individual calls per SNP, resulting in accurate genotype calls (Oliphant et al., 2002). Another advantage of SNP chips is the evaluation of the same loci across all individuals per experiment, which is possibly more difficult to achieve with GBS within experiment and across experiments. The latter is because GBS methods are based on generating sequencing libraries with restriction enzyme digested DNA that leads to variance representation of loci among individuals. Some of these limitations could be overcome by genotype imputation (Li et al., 2009) if a reference panel of genotypes is available. The latter is mostly lacking for non-model organisms.

The main purpose in using commercially available SNP chips is the identification of conserved cross species SNPs, reported in the literature as cross-species amplification, cross-amplification or cross-species genome-wide arrays. For example, Malhi et al. (2011) genotyped seven old world monkey species using an Illumina Golden Gate Array of *Macaca mulatta*, a closely related species, reporting 173 polymorphic SNPs. Likewise, Miller et al. (2012) studied the relationship between the successful applicability of cross-species SNP microarrays and evolutionary time using OvineSNP50, BovineSNP50 and EquineSNP50 BeadChips to identify SNPs in target wild species. They reported that the call rate decreased ∼1.5% per each million years of divergence time between species and the polymorphism retention of SNPs declined exponentially leveling off after about 5 Myr of divergence. Moreover, SNP genotyping in wood bison, plains bison and European bison (Pertoldi et al., 2010), scimitar-horned and Arabian oryx (Ogden et al., 2012) were performed using the Illumina BovineSNP50 BeadChip, reporting 1524, 1403, 929, 148, and 149 polymorphic SNPs, respectively. SNP genotyping in dromedary was performed using the Illumina Bovine 777K SNP BeadChip and the Illumina Ovine 600K SNP BeadChip microarrays (Bertolini et al., 2017), reporting 29900 bovine and 14179 ovine SNPs successfully genotyped.

Kharzinova et al. (2015) also reported that 43.0 and 47.0% of all SNPs in the Illumina BovineSNP50 BeadChip and the Illumina OvineSNP50 BeadChip, respectively, could be genotyped in reindeer. In addition, Haynes and Latch (2012) and Moravèíková et al. (2015) reported that 38.7 and 53.89% of the SNPs in the Illumina Bovine SNP50 BeadChip, respectively, were identified in cervids, in at least 90% of individuals, despite 25.1–30.1 million years divergence between Bovidae and Cervidae (Hassanin and Douzery, 2003). Furthermore, Hoffman et al. (2013) reported that 19.2% of all SNPs of the Illumina CanineHD BeadChip could be genotyped in seals, and reported 173 polymorphic SNPs despite a phylogenetical divergence time of around 44 million years. Therefore, the use of SNP microarrays of species with well-studied genomes have the potential to identify SNPs in related and widely diverged species.

Interestingly, all of the reported cross species analysis used different versions of GenomeStudio (Illumina, United States) and were not comparable as each research group gave different

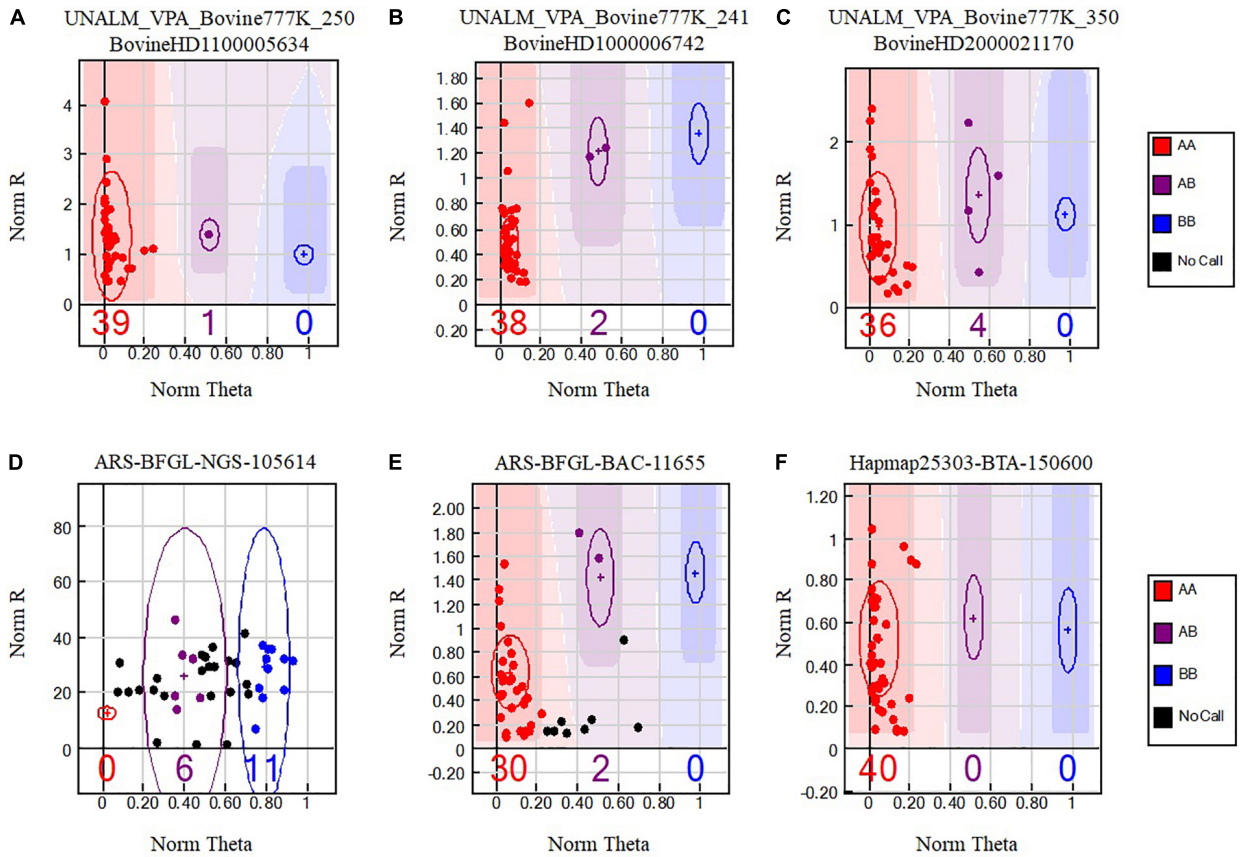


FIGURE 1 | Single nucleotide polymorphisms (SNPs) from the BovineHD Genotyping Beadchip that cross-amplify in alpacas. Genotypes are called for each sample (dot) by their signal intensity (Norm R, Y-axis) and Allele Frequency (Norm Theta, X-axis) relative to canonical cluster positions (dark shading) for a given SNP marker (red = AA, purple = AB, blue = BB). Black points represent no call samples. Polymorphic SNP genotype cluster patterns **(A–C)** selected with Method VI are presented. Difficult to interpret genotype cluster patterns **(D,E)** that were not retained are also shown and a monomorphic genotype cluster pattern **(F)** present among the 6756 positive SNPs.

weights to parameters used to generate their genotyping results. Haynes and Latch (2012) and Moravèíková et al. (2015) used a Call Frequency (Call Freq) $\geq$ 0.9 while Pertoldi et al. (2010) and Kharzinova et al. (2015) used a Call Freq = 1. Call Frequency was calculated as the number of genotype calls divided by the sum of no-calls and calls for each SNP. Lower Call-Frequency increases accuracy (Oliphant et al., 2002). Aiming at increasing the stringency of the analysis other research groups considered GenTrain score $\geq$ 0.25 (Hoffman et al., 2013) or the average GC score (average GC) $\geq$ 0.7 (Bertolini et al., 2017). The GenTrain score takes into account the quality and shape of the genotype clusters (**Figure 1**) and their relative distances from one another for each SNP while the average GC is calculated for each SNP as the average of the 10^{th} percentile and 50^{th} percentile of the distribution of GenCall scores.

Given the above experiences, the aim of this study was to evaluate SNP genotyping in alpacas using the BovineHD Genotyping Beadchip (Illumina, United States), in spite of 42.7 million years of evolutionary divergence between these two species (Wu et al., 2014) and to evaluate the different analysis methods reported in the literature.

MATERIALS AND METHODS

DNA Samples and Genotyping

Blood samples from 40 Huacaya type alpacas (4 females and 36 males) were collected by venipuncture and transferred to FTA cards. Organic DNA extraction and genotyping was done at Neogen-Geneseek laboratories (United States). Samples were genotyped using the BovineHD Genotyping Beadchip (777962 SNPs, Illumina). The sample set of unrelated animals originated from two geographical distinct Andean regions and from two separate alpaca farms within region, Chagas Chico and San Pedro de Racco in the central Andes and INCA TOPS S.A. and MICHELL & CIA S.A in the most southern Andes. The number of animals used for this study was determine to be the minimum necessary to identify SNPs with minor allele frequency (MAFs) = 0.0125 that will allow to observe at least one heterozygous genotype per sample and per SNP.

Data Analysis

Bioinformatics analysis was performed at the Universidad Nacional Agraria La Molina, Lima, Peru. The software

TABLE 1 | Parameter values used for each method of analysis.

Parameter	Method I	Method II	Method III	Method IV	Method V	Method VI	Method VII
No-call threshold	≥0.05	≥0.05	≥0.15	≥0.15	≥0.25	≥0.25	≥0.25
Call frequency	≥0.9	1	≥0.9	1	≥0.9	1	1
Average GC	≥0.7	≥0.7	≥0.7	≥0.7	≥0.7	≥0.7	*
GenTrain score	≥0.25	≥0.25	≥0.25	≥0.25	≥0.25	≥0.25	≥0.25

**In Method VII, average GC ≥ 0.7 parameter was not applied.*

GenomeStudio 2011.1 (Genotyping module version 1.9.4, Illumina, United States) was used to analyze the genotyping reports. GenomeStudio normalizes the intensities of signals for each locus and assigns a cluster position for each sample. Three parameters, no-call threshold, call frequency, and average GC were evaluated. No-call threshold or GenCall score cutoff is a quality metric calculated for each genotype (data point) and ranges from zero to one. GenCall scores decrease in value the further they are from the center of the cluster to which they are associated (**Figures 1A–C**). A no-call threshold of 0.15 is normally used for analysis of Infinium data when genotyping the same species. Hence, genotypes with GenCall scores less than 0.15 are not assigned genotypes because of being far away from the center of a cluster and therefore are categorized as a no call for the locus (**Figures 1D,E**; black dots). Call Frequency is calculated as the number of genotype calls divided by the sum of no-calls and calls for each SNP. The average GC is the simple average of the 50%GC and the 10%GC scores calculated for each SNP, where the 50%GC score represent the 50th percentile of GenCall scores across all called genotypes and the 10%GC score represents the 10th percentile. The parameters of call frequency, 50%GC and 10%GC evaluate the quality and performance of DNA samples within an experiment. Our analysis was performed using seven combinations of values for the latter three parameters. These seven combinations are labeled as Methods and are presented in **Table 1**. These methods aimed at comparing the effect of call frequency 0.9 and 1 under different no-call threshold values of ≥0.05 (Method I and Method II), ≥0.15 (Method III and Method IV) and a more stringent no-call threshold ≥0.25 (Method V, Method VI. and Method VII; Hoffman et al., 2013) in selecting SNPs. The average GC score calculated for each SNP ranks the genotype call signal from 0 (bad) to 1 (good) (Bertolini et al., 2017). We have used an average GC score value of ≥0.7 for all methods except Method VII. Similarly, a GenTrain score ≥ 0.25 (Hoffman et al., 2013) was used for all methods evaluated. The GenTrain score, calculated for each SNP by GenomeStudio, takes into account the shape of the genotype cluster and their relative distance from one another within a cluster. For all methods, positive SNPs with MAF ≥ 0.01 were retained as polymorphic SNPs.

Alignment of Flanking Sequence of Alpaca Positive Bovine SNPs With Reference Alpaca Genomes

To confirm that discovered alpaca SNPs were indeed polymorphic, two alpaca genome assemblies [Vicugna_pacos-2.0.2, GCA_000164845.3, with 22X coverage and assembled into 3374 scaffolds (KB632434-KB635807); and Vi_pacos_V1.0, GCA_000767525.1, with 72.5X coverage and assembled into 4322 scaffolds (KN266727–KN271048)] were used to align flanking sequences of alpaca positive polymorphic bovine SNPs for each method under comparison.

Microarray genotyping of SNPs result from hybridizing denatured fragments of the DNA being genotyped (target DNA) to 50 bp long SNP probes anchored on beads within a microarray chip. We hypothesize that for the identification of positive SNPs at least the first 21 to 26 nucleotides flanking the polymorphic nucleotide of the probe would need to be 100% similar to the target DNA, allowing for the rest of the probe and target sequences less than perfect similarity while permitting the priming extension of the probe fragment by the polymerase. This latter hypothesis is supported in part by Sechi et al. (2009) who reported that increased sequence divergence (mismatches) toward the 3′ end of the probe immediately flanking the variant nucleotide would have the greatest destabilizing hybridization effect resulting in no calls. Therefore, the 5′ end sequences used for BLAST analysis started at the 20th or 25th nucleotide 5′ to the polymorphic nucleotide and ended with allele A or allele B of the polymorphic nucleotide at the 3′ end. Conversely, the 3′ end flanking sequences were read on the negative DNA strand, started at the allele A or allele B of the polymorphic nucleotide, and ended at the 20th and 25th nucleotide at its 5′ end. These alignments were performed using the BLAST (blastn-short task) software of the Galaxy Platform hosted at the Minnesota Supercomputing Institute (University of Minnesota). SNPs flanking sequences that showed perfect alignments were selected, and a list with these SNPs was generated for each alpaca reference genome. Only SNPs that were unique and detected in both reference genomes were retained. Since only 100% sequence similarity between a positive bovine SNP and the alpaca genome was observed for the first 20 or 25 nucleotides flanking the variant nucleotide, the rest of the sequence to generate the 101 nucleotide sequence of alpaca SNPs was retrieve from the Vicugna_pacos 2.0.2. Hardy–Weinberg equilibrium, based on genotype distributions for each SNP, was evaluated with Genpop (Rousset, 2017) and ChiTest_p100 (Illumina Proprietary, 2008). Finally, these SNPs were assigned to alpaca chromosomes based on chromosome syntenies between cattle and camelid as described by Balmus et al. (2007) and scaffold assignments to chromosomes as described by Avila et al. (2014). Since, the phylogenetic analysis done by Kadwell et al. (2001) suggested a Latin name change for alpacas to *Vicugna pacos*; we have adopted the acronym VPA for alpaca chromosomal naming in this manuscript.

TABLE 2 | Number of positive SNPs by method.

Parameter of analysis	Method I	Method II	Method III	Method IV	Method V	Method VI	Method VII
	No-call threshold ≥0.05		No-call threshold ≥0.15		No-call threshold ≥0.25		
	Call freq ≥ 0.9	Call freq = 1	Call freq ≥ 0.9	Call freq = 1	Call freq ≥ 0.9	Call freq = 1	Call freq = 1
Call frequency	530106	111471	368001	39279	262506	23429	23429
Average GC (≥0.7)	22437	11364	24979	8232	25609	6756	*
GenTrainScore (≥0.25)	22437	11364	24979	8232	25609	6756	23429
MAF (≥0.01)	22435	11364	24962	8232	25563	6756	23427
MAF (≥0.05)	1970	898	1724	430	1467	274	2044

**In Method VII, average GC ≥ 0.7 parameter was not applied.*

Identification of Nearest Genes to Alpaca Polymorphic SNPs and Alpaca/Dromedary SNPs

The Vicugna_pacos 2.0.2 reference genome was used to identify the most proximal gene to each polymorphic SNPs. A list of these genes was develop and used for gene ontology (GO) analysis[1] for biological process GO terms. Similarly, we aligned alpaca polymorphic SNP sequences to the dromedary reference genome (PRJNA234474_Ca_dromedarius_V1.0, GCF_000767585.1) to assess SNP sequence conservation between alpaca and dromedary.

RESULTS

The number of bovine SNPs yielding positives signals are reported in **Table 2** for each of the analysis methods as described in **Table 1**. As expected, the parameters call frequency and no-call threshold had an inverse effect on the total number of positive SNPs, decreasing in number as no-call threshold and call frequency increased. Out of the 777962 SNPs analyzed 68.1, 47.3, and 33.7% were detected with a call frequency of 0.9 (Methods I, III, and V), while 14.3, 5.1, and 3.0% were detected with a call frequency of 1 (Methods II, IV, and VI, respectively). However, when average GC ≥ 0.7 was applied, a further reduction of positive SNPs was observed with 2.9, 1.5, 3.2, 1.1, 3.3, and 0.9% for Methods I, II, III, IV, V, and VI, respectively.

The percentage decrease in positive SNPs observed between Methods I and II is 21.0%, Methods III and IV is 10.7%, and Methods V and VI is 8.9%. Hence, the percentage difference of positive SNPs within a no-call threshold value decreases as the call frequency increases. However, this decrease is less pronounced as the no-call threshold increased. The differences of detected SNPs between Method I and Method II (53.8%), Method III and Method IV (42.3%) and, Method V and Method VI (30.7%), suggested that the effect of call frequency decreases when the no-call threshold increases.

The comparison of results between Methods VI and VII illustrate the effect of the average GC parameter. The number of retained SNPs in Method VI is 6756 representing a reduction of 71.2% when compared to Method VII. Hence, the effect of the average GC parameter was important in reducing the number of false positive SNPs. The GenTrain score ≥ 0.25 did not show any effect on the number of retained SNPs when the average GC ≥ 0.7 was applied. In **Supplementary Table S1** we present the minimum, maximum, mean, and standard deviation scores of average GC and GenTrain score observed for each method. However, we did not test if these latter two parameters are interchangeable.

Significant reduction in the number of SNPs retained was observed when SNPs with MAF ≥ 0.05 are selected going from 91% reduction for Method I to 96% for Method VI. Under the conditions of our analysis, Method VI showed the highest stringency and identified 6756 SNPs with MAF ≥ 0.01.

In **Table 3** we present results obtained from the alignment of all retained SNPs, with MAF ≥ 0.05, to both alpaca reference genomes. Likewise, similar analysis is presented for Method VI for SNPs with MAF ≥ 0.01.

Out of all the polymorphic SNPs with MAF ≥ 0.05 presented in **Table 2**, 5.3, 5.6, 4.6, 6.1, 5.0, 6.9 and 8.0%, were aligned to the Vicugna_pacos-2.0.2 genome assembly for Methods I, II, III, IV, V, VI, and VII, respectively. Moreover, 5.3, 5.2, 4.5, 5.6, 5.0, 6.6, and 7.7% were aligned to the Vi_pacos_V1.0 genome assembly for Methods I, II, III, IV, V, VI, and VII, respectively. Some of the SNPs with MAF ≥ 0.05 presented in **Table 2** were identified in more than one scaffold and a few were repeated within a single scaffold. Therefore, only 4.0, 4.0, 3.6, 4.2, 4.0, 5.8, and 6.3% were unique and were common to both genomes, for Methods I, II, III, IV, V, VI, and VII, respectively.

From the unique SNPs identified for each method we could only assigned 57, 29, 49, 15, 45, 13, and 98 SNPs to alpaca chromosomes for Methods I, II, III, IV, V, VI, and VII, respectively. These assignments are based on chromosome homology between cattle and camelid described by Balmus et al. (2007) or based on the cytogenetic map information developed by Avila et al. (2014).

Since the no-call threshold, call frequency, and average GC parameters were more stringent for Method VI, we selected the 400 unique SNPs with MAF ≥ 0.01 common to both reference genomes as a new set of alpaca SNPs identified in this study. The MAFs of these SNPs ranged from 0.0125 to 0.075 of which 342 SNPs had a MAF = 0.0375 (**Supplementary Table S2**) and only seven SNPs were not in Hardy–Weinberg equilibrium. In **Figure 1** we present three examples of selected unique and three unselected SNPs obtained with Method VI. All 400 SNPs showed the classical genotype cluster pattern expected from polymorphic SNPs (**Figures 1A–C**) while

[1] geneontology.org

TABLE 3 | Number of positive bovine SNPs aligned to the alpaca reference genomes.

Reference genome	Method I	Method II	Method III	Method IV	Method V	Method VI	Method VII	Method VI
				MAF ≥ 0.05				MAF ≥ 0.01
Vicugna_pacos-2.0.2								
Aligned to more than one scaffold	10	5	7	3	7	1	9	33
Unique SNPs	94	45	72	23	67	18	154	467
Vi_pacos_V1.0								
Aligned to more than one scaffold	10	6	6	3	6	1	11	30
Unique SNPs	95	41	72	21	68	17	146	466
SNPs common to both reference genomes	79	36	62	18	59	16	129	400
SNPs with predicted chromosomal localization	57	29	49	15	45	13	98	292

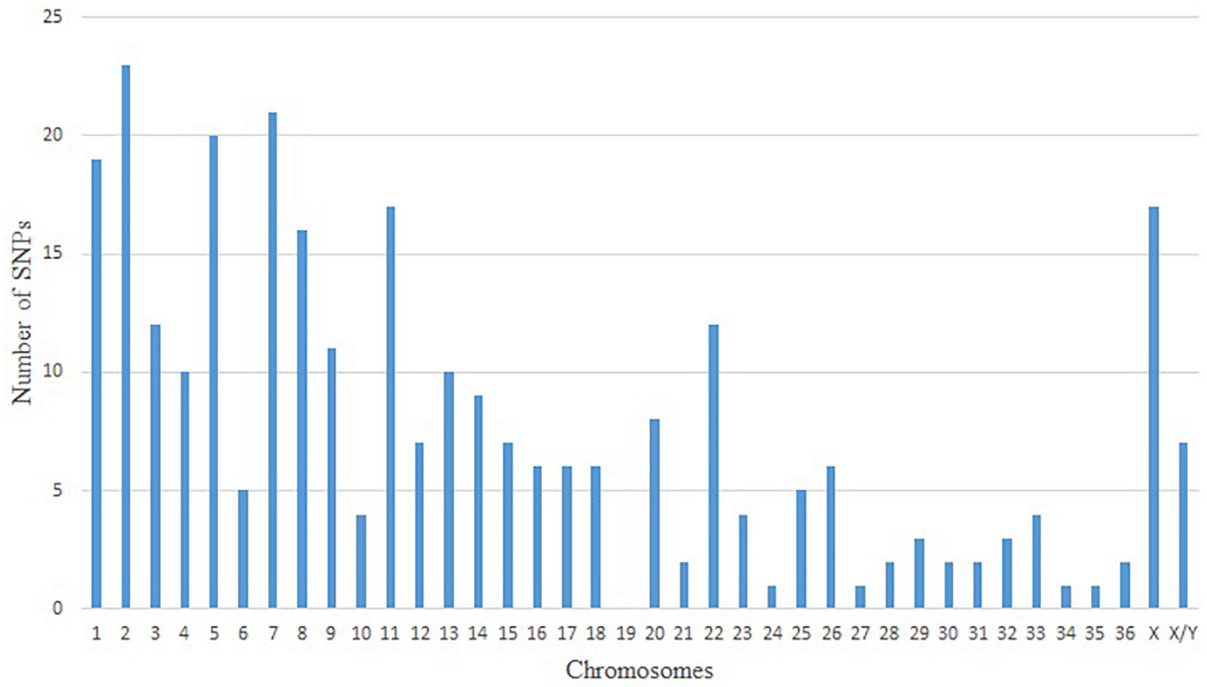


FIGURE 2 | Distribution of positive and unique SNPs in alpaca chromosomes by predicted localization.

the unselected showed difficult to interpret genotype cluster patterns (**Figures 1D,E**) with the exception of monomorphic SNPs (**Figure 1F**). Of the 400 unique SNPs, 292 SNPs were assigned to alpaca specific chromosomes (**Figure 2** and **Supplementary Table S2**). Interestingly, no SNP was assigned to VPA19.

Of the 400 polymorphic 209 were localized within 202 annotated alpaca genes (Vicugna_Pacos-2.0.2) and 69 of 400 SNPs showed perfect flanking alignment of 101 nucleotides between alpaca and dromedary. Moreover, 29 SNPs of the 69 SNPs were localized in similarly annotated dromedary and alpaca genes (**Supplementary Table S3**). The ontology analysis of the 202 annotated genes displays five GO terms that were enriched for genes at the polymorphic SNPs. The five GO terms identified were, (1) positive regulation of synaptic transmission (10 genes), (2) cell morphogenesis (20 genes), (3) cell adhesion (24 genes), (4) generation of neurons (35 genes), and (5) regulation of multicellular processes (52 genes). The majority of these genes are involved in biological developmental processes.

DISCUSSION

The application of genome wide association studies (GWASs) studies to alpacas will only be possible when enough SNPs are identified to provide a reasonable coverage of their genome. This study tested a cross hybridization approach for the identification of conserved polymorphic cattle/alpaca SNPs using the available BovineHD Genotyping Beadchip. The assessment of combination of scores for no-call threshold, call frequency and average GC yielded an optimum method that identified 400 conserved polymorphic SNPs. However, these latter SNPs are affected by ascertainment bias because of our small sample population and lack of information as to whether the SNPs originate from coding or non-coding regions that influence their minor allele

frequencies. This small sample population will allow to detect SNPs with MAF $\geq$ 0.0125 therefore rare SNPs will not be represented. It has been suggested by Hoffman et al. (2013) that SNPs cross-amplified from high-density arrays might be enriched for conserved genomic regions retaining ancestral polymorphisms. However, the commercially available Bovine HD SNP chip we used in this study was designed to provide uniform genome coverage with evenly spaced SNPs and therefore it can be inferred that our discovered SNPs are selectively neutral. Nielsen (2004) provides a thorough review on ascertainment bias for SNP data.

One measure of genotyping success is the SNP conversion rate defined as the proportion of all genotyped SNPs showing clear genotyping clusters (**Figures 1A–C**). Our conversion rate was very low (0.008%) and is in line with observed conversion rates for cross hybridization genotyping experiments (Hoffman et al., 2013). The 400 polymorphic SNPs remain to be validated by genotyping a different and larger alpaca population sample.

Data Analysis

The percentage of SNPs identified in at least 90% of samples by Method I and Method III was higher, 68.1 and 47.3%, respectively, than those SNPs found in the genotyping of deers (38.7%, Haynes and Latch, 2012). Moreover, the percentage of SNPs in Method I was also higher than those SNPs reported in cervids (53.9%, Moravèíková et al., 2015) using the Illumina BovineSNP50 Bead Chip. However, the percentage of SNPs observed using a more stringent no-call threshold (Method III) was less than these reports.

The percentages of SNPs identified with call frequency = 1 in Methods II (14.3%), IV (5.1%), and VI (3.0%) were less than those found in the genotyping of bisons (97.0%, Pertoldi et al., 2010) using the Illumina BovineSNP50 Bead Chip, and reindeers (43.0%, Kharzinova et al., 2015) using the Illumina BovineSNP50 v2. Bead Chip.

The percentages of SNPs identified with call frequency = 1 and selected based on their average GC $\geq$ 0.7 in Methods II (1.5%), IV (1.0%), and VI (0.9%) are less than those found in the genotyping of camels (3.8%, Bertolini et al., 2017) using the Illumina Bovine 777K SNP BeadChip. This could be due to higher heterogeneity of the dromedary sample in comparison to our alpaca sample set and/or in part determined by higher false positives identified in the dromedary-bovine cross hybridization experiments as stated by Bertolini et al. (2017).

The effects of the no-call threshold $\geq$ 0.25, call frequency = 1 and average GC $\geq$ 0.7 were significant in reducing the number of positive SNPs. However, under the conditions imposed by our analysis the use of GenTrain score threshold $\geq$ 0.25 (Hoffman et al., 2013) did not have any effect on the identification of positive SNPs in all methods at an average GC $\geq$ 0.7. However, it cannot be discarded that the GenTrain score threshold $\geq$ 0.25 might have a similar effect if it is used in substitution of the average GC $\geq$ 0.7 parameter.

The percentage of polymorphic SNPs in Methods II (1.5%), IV (1.1%), and VI (0.9%) is less than those found in the genotyping of deers (2%, Haynes and Latch, 2012), bisons (4.1%, Pertoldi et al., 2010), cervids (2.8%, Moravèíková et al., 2015), reindeers (2.3%, Kharzinova et al., 2015), and camels (3.6%, Bertolini et al., 2017). When a call frequency of 0.9 was used [Methods I (2.3%), III (3.2%), and V (3.3%)], the percentage of retained SNPs was higher in comparison to those reported by Haynes and Latch (2012); Kharzinova et al. (2015), and Moravèíková et al. (2015). In addition, the number of SNPs with MAF $\geq$ 0.05 were rare among the 40 samples analyzed.

Method VI identified 6756 SNPs with MAF $\geq$ 0.01 of which 400 showed perfect flanking alignment of 20 or 25 nucleotides adjacent to the polymorphic nucleotide and were further analyzed by manually observing their genotype cluster distributions where at least one sample was identified as heterozygous for each SNP. When applying the exponential polymorphic decay function developed by Miller et al. (2012) to our findings, the expected percentage of polymorphic SNPs is 0.000515% and our observed 6756 SNPs with MAF $\geq$ 0.01 identified with Method VI represent 0.008684%, which is 16.5 times higher than expected. However, this observed number of SNPs could represent an overestimate since we have not ascertained the polymorphic status of each of these putative SNPs. However, the 400 polymorphic SNPs reported in this study represent 0.000514%, which is similar to the calculated expected percentage of polymorphic SNPs obtained with the exponential decay function formula developed by Miller et al. (2012). Examples of polymorphic SNPs discovered in this study are presented in **Figures 1A–C**, showing the genotype cluster distributions of positively identified SNPs. For illustration purposes, we also present cluster distributions of two SNPs that are difficult to interpret and were not retained (**Figures 1D,E**) with our analysis as well as a monomorphic SNP (**Figure 1F**). The so-called monomorphic SNPs, represent alpaca DNA fragments that have hybridized to specific probes in the SNP chip and are homozygous for the A or the B alleles in the sample population. These monomorphic SNPs could also be referred as false negatives. Monomorphic SNPs could very well be polymorphic SNPs if a larger sample set or a different sample set is used.

Only 292 out of the 400 polymorphic SNPs were mapped to alpaca chromosomes and 108 (27%) could not be assigned to chromosomes with available indirect methods (Balmus et al., 2007; Avila et al., 2014). The absence of SNPs assigned to VPA19 and the low number of SNPs ($\geq$5) assigned to 14 other chromosomes is difficult to explained with our available data. In this study, all SNPs identified using Method VI were located across all bovine chromosomes (**Supplementary Figure S1**). Bertolini et al. (2017) also reported this latter distribution for dromedary SNPs. In this study, of SNPs identified by less stringency methods (Method I and Method III) localized one bovine SNP (BovineHD1300018765) on VPA19. Hence, we believe that the observed distribution of SNPs across chromosomes is due to the stringency applied in Method VI and our inability to chromosomally assigned 27% of the identified SNPs based on the level of resolution of the methods used, in this study, to infer alpaca chromosomal assignments.

A comparison of the 400 SNP sequences between alpaca and dromedary identified 209 of the 400 SNPs to be localized within 202 annotated alpaca genes (Vicugna_Pacos-2.0.2) and 69 SNPs showed perfect flanking alignment of 101 nucleotides between alpaca (Vicugna_Pacos-2.0.2) and dromedary

(PRJNA234474_Ca_dromedarius_V1.0, GCF_000767585.1). Moreover, 29 SNPs out of the 69 SNPs were localized in similarly annotated dromedary and alpaca genes (**Supplementary Table S3**). An ontology analysis of the 202 annotated gene display five GO terms were identified as enriched for genes at the polymorphic SNPs that were Bonferroni corrected for $P < 0.05$. The five GO terms identified were positive regulation of synaptic transmission (10 genes), cell morphogenesis (20 genes), cell adhesion (24 genes), generation of neurons (35 genes), and regulation of multicellular processes (52 genes). The majority of these genes are involved in biological developmental processes. It is possible that for this latter reason they exhibit sequence conservation between alpaca and bovine that would explain the conserved retention of polymorphic SNPs at these loci. However, because of our small sample size and small number of genes associated to polymorphic SNPs, the latter analysis should be treated with caution.

CONCLUSION

In spite of 42.7 million years of evolutionary divergence between cattle and alpacas (Wu et al., 2014), the application of the cross hybridization approach for the identification of polymorphic alpaca SNPs, based on the use of the BovineHD Genotyping Beadchip (Illumina), was successful. The comparison of different filtering methods indicated that no-call threshold, call frequency and average GC are important parameters to consider for the successful identification of polymorphic SNPs in cross hybridization experiments. Based on our results, the filters of no call threshold ≥ 0.25, call frequency = 1, average GC ≥ 0.7, and GenTrain score ≥ 0.25 are recommend for detection of SNPs in non-model species. The application of these filters allowed the identification of 6756 alpaca SNPs of which 400 are polymorphic and 292 SNPs were assigned to alpaca chromosomes. Further, 209 SNP were localized in 202 alpaca gene sequences and 29 of these were also located at similar gene loci in dromedary. Of the 400 alpaca SNPs, 69 shared 100% percent sequence similarity to dromedary. Our results represent a significant increase in polymorphic molecular markers for alpaca at this moment and indicates that investing in discovering SNPs by GBS or by sequencing reduced representation libraries of a larger number of samples would be necessary to generate an alpaca SNP chip for the successful application of GWAS to this species.

ETHICS STATEMENT

The Universidad Nacional Agraria La Molina has recently established an Ethics Committee for Scientific Research by University Resolution No. 0345-2018-CU-UNALM of October 22, 2018 which has not initiated its operations as of yet. However, we have a letter signed by the Dean of the college of Animal Sciences corroborating that the protocol used for blood collection titled "Collection of Blood for FTA cards" is of conventional application and it follows the requirements of the National Act No. 30407 "Ley de Proteccion y Bienestar Animal" (Act for the Protection and Well-being of Animals).

AUTHOR CONTRIBUTIONS

FPL and MR conceived the study. MM, GG, and FPL participated in data analysis. MM and FPL co-wrote the manuscript. GG and FPL supervised the study. MR and FB reviewed and corrected the manuscript. All authors read and approved the manuscript.

FUNDING

The authors acknowledge the financial support from CONCYTEC through project 125-2015 FONDECYT, and VLIR-UOS funding to the UNALM (IUC) programme. Opinions of the author(s) do not automatically reflect those of either the Belgian government or VLIR-UOS, and can bind neither the Belgian Government nor VLIR-UOS. Funding was also provided, in part, by Hatch project MIN-16-103, MN Experiment Station, the State of Iowa and the Ensminger Endowment Fund.

ACKNOWLEDGMENTS

The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper (http://www.msi.umn.edu). Likewise, authors acknowledge the farm communities of Chagas Chico and San Pedro de Racco and, INCA TOPS S.A. and MICHELL & CIA S.A. for facilitating the collection of alpaca blood samples at their facilities. They are grateful to the reviewers for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00361/full#supplementary-material

REFERENCES

Avila, F., Baily, M., Perelman, P., Das, P., Pontius, J., Chowdhary, R., et al. (2014). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*Lama pacos*). *Cytogenet. Genome Res.* 144, 193–204. doi: 10.1159/000370329

Balmus, G., Trifonov, V., Biltueva, L., O'brien, P., Alkalaeva, E., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative Cetartiodactyla ancestral karyotype. *Chromosome Res.* 15, 499–515. doi: 10.1007/s10577-007-1154-x

Bertolini, F., Elbeltagy, A., and Rothschild, M. (2017). Evaluation of the application of bovine, ovine and caprine SNP chips to dromedary genotyping. *Livestock Res. Rural Dev.* 29:31.

Chandramohan, B., Renieri, C., La Manna, V., and La Terza, A. (2015). The alpaca melanocortin 1 receptor: gene mutations, transcripts, and relative levels of

expression in ventral skin biopsies. *Sci. World J.* 2015:265751. doi: 10.1155/2015/265751

Dekkers, J. (2012). Application of genomics tools to animal breeding. *Curr. Genomics* 13, 207–212. doi: 10.2174/138920212800543057

Guridi, M., Soret, B., Alfonso, L., and Arana, A. (2011). Single nucleotide polymorphisms in the melanocortin 1 receptor gene are linked with lightness of fibre colour in Peruvian Alpaca (*Vicugna pacos*). *Anim. Genet.* 42, 679–682. doi: 10.1111/j.1365-2052.2011.02205.x

Hassanin, A., and Douzery, E. (2003). Molecular and morphological phylogenies of ruminantia and the alternative position of the moschidae. *Syst. Biol.* 52, 206–228. doi: 10.1080/10635150390192726

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53, 876–883. doi: 10.1139/G10-076

Haynes, G., and Latch, E. (2012). Identification of novel single nucleotide polymorphisms (SNPs) in Deer (*Odocoileus* spp.) using the BovineSNP50 BeadChip. *PLoS One* 7:e36536. doi: 10.1371/journal.pone.0036536

Hoffman, J., Thorne, M., McEwing, R., Forcada, J., and Ogden, R. (2013). Cross-amplification and validation of SNPs conserved over 44 million years between seals and dogs. *PLoS One* 8:e68365. doi: 10.1371/journal.pone.0068365

Illumina Proprietary (2008). *GenomeStudioTM Genotyping Module v1.0 User Guide*. Available at: https://www.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/genomestudio/genomestudio-2011-1/genomestudio-gt-module-v1-0-user-guide-11319113-a.pdf (accessed January 7, 2019).

Kadwell, M., Fernandez, M., Kadwell, H., and Baldi, R. (2001). Genetic analysis reveals the wild ancestors of the llama and the alpaca. *Proc. R. Soc. Lond.* 268, 2575–2584.

Kharzinova, V., Sermyagin, A., Gladyr, E., Okhlopkov, I., Brem, G., and Zinovieva, N. (2015). A study of applicability of SNP chips developed for bovine and ovine species to whole-genome analysis of reindeer *Rangifer tarandus*. *J. Hered.* 106, 758–761. doi: 10.1093/jhered/esv081

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. *Annu. Rev. Genomics Hum. Genet.* 10, 387–406. doi: 10.1146/annurev.genom.9.081307.164242

Malhi, R., Satkoski, J., Shattuck, M., Johnson, J., Chakraborty, D., Kanthaswamy, S., et al. (2011). Genotyping single nucleotide polymorphisms (SNPs) across species in old world monkeys. *Am. J. Primatol.* 73, 1031–1040. doi: 10.1002/ajp.20969

Miller, J., Kijas, J., Heaton, M., McEwan, J., and Coltman, D. (2012). Consistent divergence times and allele sharing measured from cross-species application of SNP chips developed for three domestic species. *Mol. Ecol. Res.* 12, 1145–1150. doi: 10.1111/1755-0998.12017

Ministerio de Agricultura y Riego [MINAGRI] (2017). *Diagnostico de crianzas priorizadas para el. Plan Ganadero 2017-2012*. Lima: Autor.

Moravèíková, N., Kirchner, R., Šidlová, V., Kasarda, R., and Trakovická, A. (2015). Estimation of genomic variation in cervids using cross-species application of SNP arrays. *Poljoprivreda* 21(Suppl. 6), 33–36. doi: 10.18047/poljo.21.1.sup.6

Munyard, K., Ledger, J., Lee, C., Babra, C., and Groth, D. (2009). Characterization and multiplex genotyping of alpaca tetranucleotide microsatellite markers. *Small Rumin. Res.* 85, 153–156. doi: 10.1016/j.smallrumres.2009.07.012

Nielsen, R. (2004). Population genetic analysis of ascertained SNP data. *Hum. Genomics* 1, 218–224.

Ogden, R., Baird, J., Senn, H., and McEwing, R. (2012). The use of cross-species genome-wide arrays to discover SNP markers for conservation genetics: a case study from Arabian and scimitar-horned oryx. *Conserv. Genet. Resour.* 4, 471–473. doi: 10.1007/s12686-011-9577-2

Oliphant, A., Barker, D., Stuelpnagel, J., and Chee, M. (2002). BeadArray technology: enabling an accurate, cost-effective approach to highthroughput genotyping. *BioTechniques* 32(Suppl. 56), 60–61.

Paredes, M., Membrillo, A., Gutiérrez, J., Cervantes, I., Azor, P., Morante, R., et al. (2014). Association of microsatellite markers with fiber diameter trait in Peruvian Alpacas (*Vicugna pacos*). *Livestock Sci.* 161, 6–16. doi: 10.1016/j.livsci.2013.12.008

Pérez-Cabal, M., Cervantes, I., Morante, R., Burgos, A., Goyache, F., and Gutiérrez, J. (2010). Analysis of the existence of major genes affecting alpaca fiber traits. *J. Anim. Sci.* 88, 3783–3788. doi: 10.2527/jas.2010-2865

Pertoldi, C., Wójcik, J., Tokarska, M., Kawałko, A., Kristensen, T., Loeschcke, V., et al. (2010). Genome variability in European and American bison detected using BovineSNP50 BeadChip. *Conserv. Genet.* 11, 627–634. doi: 10.1007/s10592-009-9977-y

Rousset, F. (2017). *Genepop Version 4.7.0*. Available at: https://kimura.univ-montp2.fr~rousset/Genepop4.7.pdf (accessed April 10, 2017).

Sechi, T., Coltman, D. W., and Kijas, J. W. (2009). Evaluation of 16 loci to examine the cross-species utility of single nucleotide polymorphism arrays. *Anim. Genet.* 41, 199–202.

Slate, J., Gratten, J., Beraldi, D., Stapley, J., Hale, M., and Pemberton, J. (2009). Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica* 136, 97–107. doi: 10.1007/s10709-008-9317-z

Warren, W., Wilson, R., and Merriwether, A. (2013). *Direct Submission. Sequence Assembly Submitted by The Genome Institute*. St. Louis, MO: Washington University School of Medicine.

Wu, H., Guang, X., and Wang, J. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 More, Gutiérrez, Rothschild, Bertolini and Ponce de León. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 May 2019
doi: 10.3389/fgene.2019.00423

A First Y-Chromosomal Haplotype Network to Investigate Male-Driven Population Dynamics in Domestic and Wild Bactrian Camels

*Sabine Felkel[1,2], Barbara Wallner[1], Battsesteg Chuluunbat[3], Adiya Yadamsuren[4,5], Bernard Faye[6], Gottfried Brem[1], Chris Walzer[7,8], on behalf of the International Camel Consortium[†] and Pamela A. Burger[7]**

[1] Institute of Animal Breeding and Genetics, University of Veterinary Medicine Vienna, Vienna, Austria, [2] Vienna Graduate School of Population Genetics, Vienna, Austria, [3] Laboratory of Genetics, Institute of Biology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia, [4] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, [5] Wild Camel Protection Foundation Mongolia, Ulaanbaatar, Mongolia, [6] CIRAD-ES, UMR 112, Campus International de Baillarguet, Montpellier, France, [7] Research Institute of Wildlife Ecology, Department of Integrative Biology and Evolution, Vetmeduni Vienna, Vienna, Austria, [8] Wildlife Conservation Society, Wildlife Health Program, Bronx, NY, United States

OPEN ACCESS

Edited by:
Fulvio Cruciani,
Sapienza University of Rome, Italy

Reviewed by:
He Meng,
Shanghai Jiao Tong University, China
Fernando Luis Mendez,
Helix OpCo LLC, United States
Ruihua Dang,
Northwest A&F University, China

***Correspondence:**
Pamela A. Burger
pamela.burger@vetmeduni.ac.at

[†]International Camel Consortium for Genetic Improvement and Conservation,
www.icc-gic.weebly.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 21 December 2018
Accepted: 17 April 2019
Published: 21 May 2019

Citation:
Felkel S, Wallner B, Chuluunbat B, Yadamsuren A, Faye B, Brem G, Walzer C and Burger PA (2019) A First Y-Chromosomal Haplotype Network to Investigate Male-Driven Population Dynamics in Domestic and Wild Bactrian Camels. Front. Genet. 10:423. doi: 10.3389/fgene.2019.00423

Polymorphic markers on the male-specific part of the Y chromosome (MSY) provide useful information for tracking male genealogies. While maternal lineages are well studied in Old World camelids using mitochondrial DNA, the lack of a Y-chromosomal reference sequence hampers the analysis of male-driven demographics. Recently, a shotgun assembly of the horse MSY was generated based on short read next generation sequencing data. The haplotype network resulting from single copy MSY variants using the assembly as a reference revealed sufficient resolution to trace individual male lines in this species. In a similar approach we generated a 3.8 Mbp sized assembly of the MSY of *Camelus bactrianus*. The camel MSY assembly was used as a reference for variant calling using short read data from eight Old World camelid individuals. Based on 596 single nucleotide variants we revealed a Y-phylogenetic network with seven haplotypes. Wild and domestic Bactrian camels were clearly separated into two different haplogroups with an estimated divergence time of 26,999 ± 2,268 years. Unexpectedly, one wild camel clustered into the domestic Bactrian camels' haplogroup. The observation of a domestic paternal lineage within the wild camel population is concerning in view of the importance to conserve the genetic integrity of these highly endangered species in their natural habitat.

Keywords: old world camelids, paternal lineage, Y chromosome, haplotype, diversity, conservation

INTRODUCTION

The male-specific region (MSY) of the mammalian Y chromosome is transferred directly from the father to the son without recombination. This unique mode of inheritance highlights the MSY as an ideal marker to study male genealogies alongside and in comparison to maternal phylogenies based on the mitochondrial DNA (mtDNA). Due to the lack of recombination, allelic states of

single nucleotide variants (SNVs) on the MSY can be combined into haplotypes and robust MSY haplotype phylogenies can be built using the principle of maximum parsimony (reviewed in Jobling and Tyler-Smith, 2017). Such MSY backbone phylogenies based on biallelic markers (Skaletsky et al., 2003) became recently available in dogs (Oetjens et al., 2018), cattle (Chen H. et al., 2018) and horses (Wallner et al., 2017; Felkel et al., 2018). The MSY revealed new insights into the domestication and uncovered male mediated historic radiations in these species. However, the Y chromosome is often not considered in genome assemblies because of the challenges to assemble its repetitive content. Hence, re-sequencing approaches to study paternal lineages often need to start with assembly work. We and others have recently shown the value that *de novo* assemblies of the MSY, generated from short read next generation sequencing (NGS) data, had for tracing male lineages (Wallner et al., 2017). We also described in detail the need to accurately define single copy regions in the assembly (Felkel et al., 2019) to unambiguously call variants in these regions only. With this approach (*de novo* assembly and SNV calling in single copy MSY regions) we elevated the resolution and accuracy of paternal lineage tracing in horses to a level similar to that in humans (Wallner et al., 2017; Felkel et al., 2019), where MSY haplotype (HT) trajectories are well described (Jobling and Tyler-Smith, 2017).

In Old World camels, domestication and historic demography have been investigated mainly by mtDNA. Domestication of dromedaries (*Camelus dromedarius*) took first place at the east coast of the Arabian Peninsula around 3,000 to 4,000 years ago (ya) (Almathen et al., 2016). Bactrian camels were domesticated around 4,000 to 6,000 ya. The described long-term divergence of 1.5 to two million years between wild (*Camelus ferus*) and domestic (*Camelus bactrianus*) two-humped camel mtDNA lineages (Mohandesan et al., 2017) predates the timeframe of camel domestication by far. Over the last century, the critically endangered wild two-humped camels (Hare, 2008) have been reduced to only ∼2,000 individuals restricted to the Mongolian Great Gobi strictly protected area "A" (GGSPAA) (Yadamsuren et al., 2012) and to the Chinese deserts Taklamakan and Lop Noor (Lei et al., 2012), where they are stringently protected. However, habitat decline, poaching, environmental pollution due to illegal mining and hybridisation with domestic Bactrian camels continue threatening the wild camels. Severely reduced mtDNA haplotype diversity (Silbermayr et al., 2010; Mohandesan et al., 2017) and low genome-wide nucleotide diversity (Fitak et al., 2016b) have been observed in the wild populations.

To trace the paternal lineages and reconstruct male-driven population dynamics in Old World camels, specifically in the highly endangered wild camel, we adopted the approach recently described in horses (Wallner et al., 2017; Felkel et al., 2019) and *de novo* assembled the Bactrian camel MSY from short read NGS data. The assembly was used as a reference for mapping and single copy MSY variant calling using short read data from five wild and three domestic Bactrian camels one dromedary as outgroup. With this, we created a first backbone Y phylogeny for Old World camels based on slowly evolving biallelic markers.

MATERIALS AND METHODS

Data Origin

Trimmed Illumina short reads from nine wild (five male/four female) and five domestic Bactrian camel (three male/two female) genomes were derived from an ongoing project (FWF P24706-B25; PI: PB) to detect signatures of selection related to domestication in Old World camelids. In addition, one male dromedary was included as outgroup. Sample details are provided in **Supplementary Table S1** and in the ethics statement below.

MSY *de novo* Assembly

Short reads of the deepest sequenced domestic male (sample ID DC269 in **Supplementary Table S1**) were mapped to the female Bactrian camel reference genome (JARL01; SAMN02053968; Jirimutu et al., 2012) using bwa aln (Li and Durbin, 2009). Unmapped read pairs were collected and converted to fastq files with samtools (Li et al., 2009), bedtools (Quinlan and Hall, 2010) and Unix commands. The resulting read pairs were *de novo* assembled with SPAdes (Bankevich et al., 2012). A schematic overview of the analytic steps is presented in **Figure 1**. For a detailed bioinformatic protocol see **Supplementary Methods**.

Classification of Single Copy MSY Regions

We mapped the Illumina short reads from the eight male and six female Bactrian camels to the raw assembly output using bwa aln, removed PCR duplicates and filtered for mapping quality > 20 using samtools. We next splitted the assembly into 50 bp windows and followed the probabilistic approach of Felkel et al. (2019), which uses the differences of normalized mapping coverages in males and females to classify the windows into single copy (scY) and multi copy (mcY) MSY and not MSY (nonMSY). For the classification we did not consider DC269 on which the reference is based. For details see **Supplementary Methods** and **Supplementary Figures S1–S5**. Contigs shorter than 200 bp and with a scY/mcY-specific content less than 50% were discarded, resulting in clean and classified MSY reference contigs.

MSY Variant Calling, Haplotype Identification and SNV Validation

Variant calling with GenomeAnalysisTK HaplotypeCaller and CombineGVCFs was performed using the mappings from above. Only variants called in scY regions (**Supplementary Table S2**), were used for downstream analyses. We excluded reference errors, phased variants, insertions and deletions (indels) and variants with multiple alternatives, heterozygous or empty calls from further analyses. A read depth of at least three in one individual and a genotype quality higher nine were set as limit to keep the variant in the list. The resulting variants were used to generate a haplotype network with Network (Bandelt et al., 1999). Missing variants (in low coverage samples or due to ambiguous mappings) have been implemented according to the samples' clustering in the network. Finally, we performed independent validation of nine SNVs via Sanger

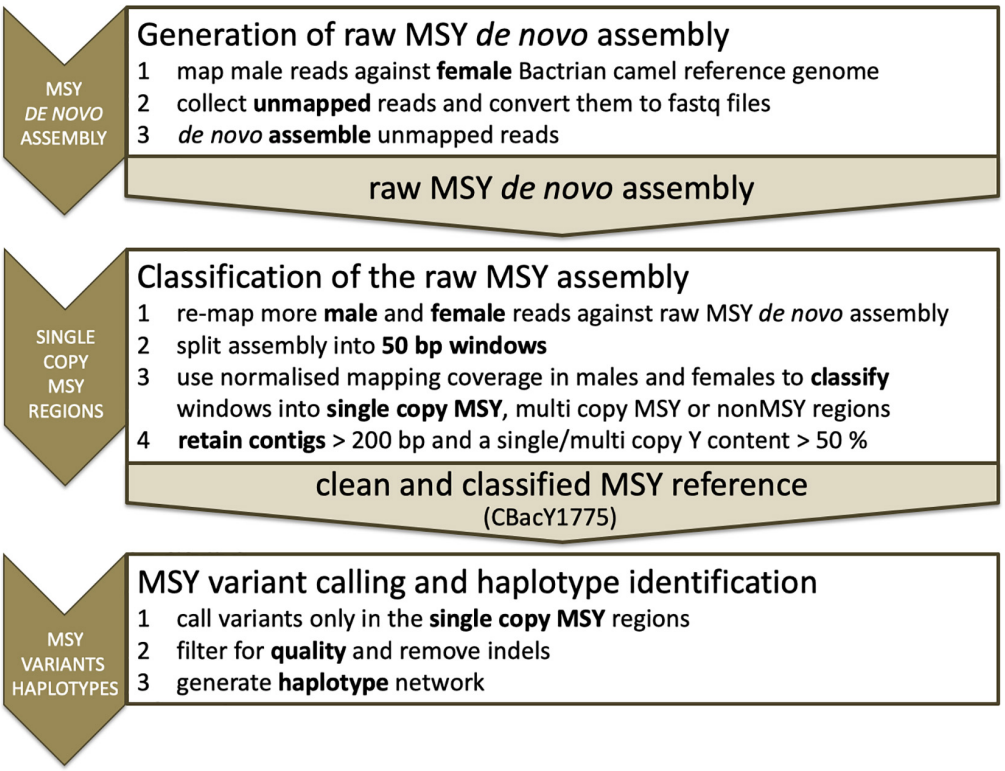


FIGURE 1 | Schematic overview of the analytic steps.

sequencing. Validation candidates and primer information are shown in **Supplementary Table S3**. Further details are given in **Supplementary Methods**.

Diversity Estimates of the Bactrian Camel MSY

Nucleotide diversity (π and θ) was calculated in R for different scenarios using the whole set of SNVs detected in all individuals but the outgroup (numbers shown in **Figure 2**).

Divergence Time Estimates Between Wild and Domestic Bactrian Camel Male Lineages

Felkel et al. (2019) calculated a mutation rate of 1.68×10^{-8} mutations/site/generation for the horse Y based on deep pedigrees and this rate is highly similar to the genome-wide estimate in humans (Kong et al., 2012). Since no pedigrees have been available to calculate a camel specific mutation rate, we used the horse rate to date nodes in the resulting camel Y network using rho statistics implemented in Network (for details see **Supplementary Methods**; variants and contigs used for dating are indicated in **Supplementary Tables S3**, **S4**, respectively; the haplotype network based on variants used for dating is shown in **Supplementary Figure S6**).

RESULTS

Bactrian Camel MSY *de novo* Assembly

We used 11,987,466 read pairs not mapping to the female Bactrian camel reference genome as input for the *de novo* assembly and classified the resulting 22,398 contigs of the raw assembly into scY, mcY, and nonMSY regions. An overview of the assembly and classification statistics is given in **Table 1**. We kept contigs longer than 200 bp and with a Y-specific content > 50% (**Supplementary Figure S5**) and retrieved the final MSY assembly (CBacY1775) with a total size of 3.8 Mbp distributed over 1,775 contigs. CBacY1775 has a minimum/maximum length of 201/87,065 bp and an N50 of 858 bp. We classified 2.39 Mbp as scY, 1.21 Mbp as mcY and 0.18 Mbp as nonMSY, respectively. scY and mcY regions are provided in **Supplementary Table S2** and the assembly can be downloaded from NCBI (RYZT00000000). Contig lengths are provided in **Supplementary Table S4**. Mappings are uploaded to the NCBI public database (Accession numbers are provided in **Supplementary Table S1**).

MSY Variation in Wild and Domestic Bactrian Camels to Build a Phylogenetic Network

For MSY haplotype reconstructions we considered only variants called in scY regions of three domestic and five wild two-humped camels. Based on a total of 596 variants

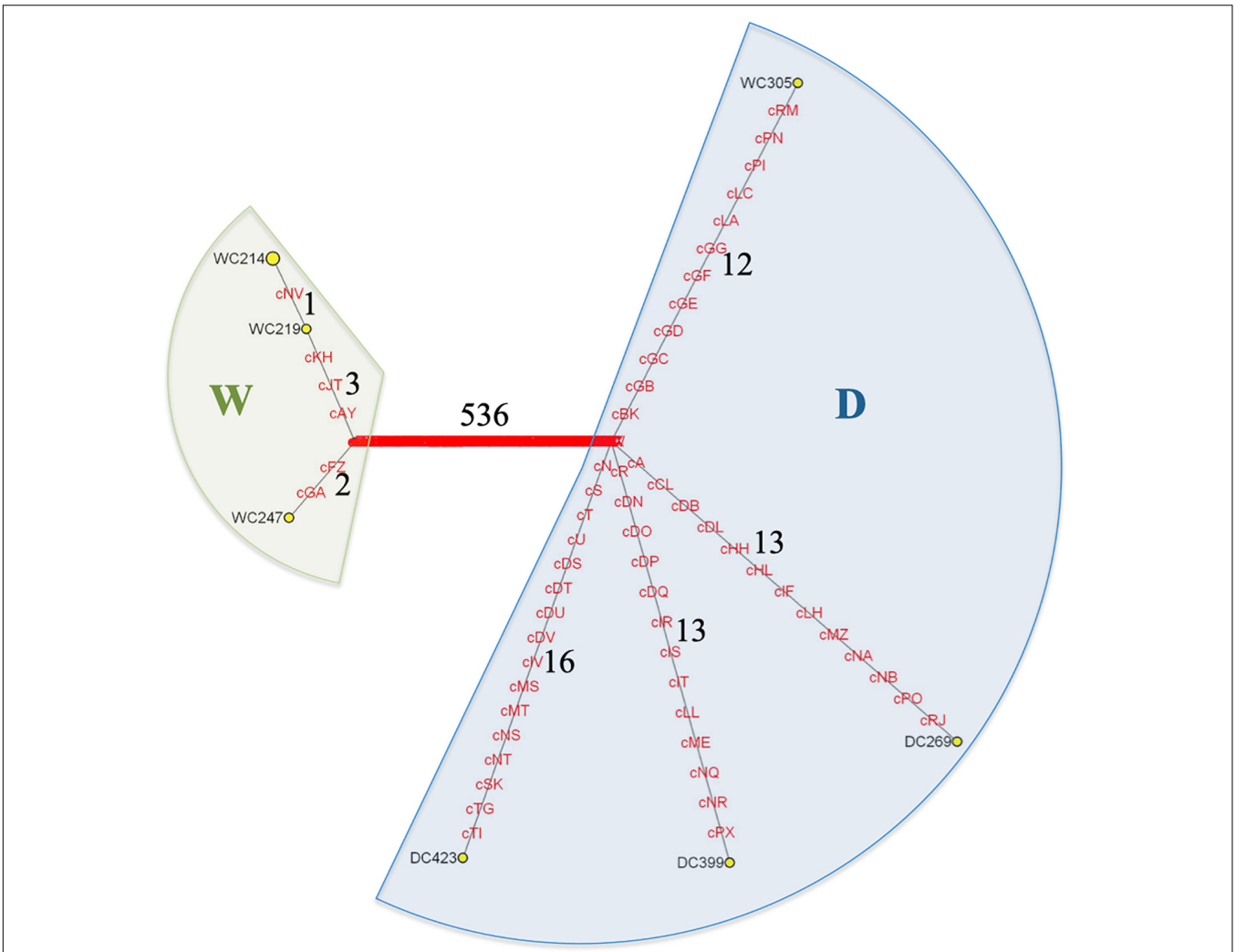


FIGURE 2 | MSY haplotype network based on 596 scY SNVs found on CBacY1775 for three domestic (DC) and five wild (WC) Bactrian camels. The outgroup is not shown. HG W (green shade) is clearly separated from HG D (blue shade). Each sample has a private haplotype except for WC214 and WC218 (circle size of WC214 corresponds to both these samples comprising this certain haplotype), which share the same father (Hare, pers. comm.). WC305 is the only wild camel belonging to HG D. Numbers indicate branch lengths (numbers of mutations). The variants correspond to those in **Supplementary Table S3**.

(**Supplementary Table S3**) we distinguished seven haplotypes that formed two clearly separated haplogroups (HG), W and D. When displayed in a phylogenetic network (**Figure 2**), HG W was private for the wild camels in our dataset, whereas HG D was detected in three domestic individuals plus one wild individual (WC305). Within HG D, the domestic Bactrian camels formed three branches, each of them with a length of 13–16 SNVs. The wild camel WC305, which unexpectedly clustered into the principally domestic HG D, had a branch length similar to the three domestic lineages (twelve SNVs). The wild camels in HG W had only two lineages with two to four variants from the basal node of the HG. Based on all detected variants we obtained diversity estimates given in **Table 2**.

We verified nine randomly selected SNV using PCR and Sanger sequencing. Six SNVs showed the expected MSY specific amplification pattern and a PCR amplicon was revealed only from male template DNA, whereas no product was generated when using female Bactrian camel DNA as template. The SNVs in these six loci were confirmed with Sanger sequencing (**Supplementary Table S3**, **Supplementary Methods**, and **Supplementary Figure S7**). However, three primer pairs developed for SNV loci (**Supplementary Table S3**) also amplified DNA in females. Although showing a single band on an agarose gel in males and females, these products could be sequenced only for one SNV (cGA). The other two SNVs could not be validated as the interferences observed in the Sanger sequence electropherograms indicated a cross amplification of MSY, autosomal and/or X-chromosomal regions. Alternative target-specific primers need to be developed for these two loci to amplify the MSY region unambiguously. Such cross amplifications are not unexpected, keeping in mind the genomic structure of the MSY (Jobling and Tyler-Smith, 2003; Skinner et al., 2013)

TABLE 1 | *De novo* assembly and classification statistics of the Bactrian camel MSY.

	Total length [Mbp]	Number of contigs	N50 [bp]	scY [Mbp]	mcY [Mbp]	nonY [Mbp]
Raw assembly	14.49	22,398	308	2.47	1.50	10.52
CBacY1775	3.79	1,775	878	2.39	1.21	0.18

with varying, often substantial sequence homology to X and autosomal regions.

Estimation of the Divergence Time Between Wild and Domestic Bactrian Camel Paternal Lineages

Using the horse Y mutation rate previously obtained and assuming a generation time of 6 years in Bactrian camels, the most recent common ancestor (MRCA) of haplogroups D and W was estimated to 26,999 ± 2,268 years before present (ybp; **Supplementary Figure S6**). The MRCAs of HG D and W were estimated to 1,764 ± 416 ybp and 490 ± 353 ybp, respectively.

DISCUSSION

De novo Assembly of the MSY Region in Bactrian Camels

In this study, we generated a 3.8 Mbp sized *de novo* assembly of the MSY region of domestic Bactrian camels and classified scY regions following the very conservative approach by Felkel et al. (2019). We then created a first Y phylogeny of highly endangered wild Bactrian camels and their domestic congeners, using a dromedary as outgroup. The size of the newly assembled MSY in Bactrian camels (3.8 Mbp) is smaller than that of horses (6.58 Mbp; Felkel et al., 2019). Apart from one male *Camelus ferus* shotgun sequenced genome (Jirimutu et al., 2012), currently available camel reference genomes were generated from female data (Wu et al., 2014; Fitak et al., 2016a; Elbers et al., 2019), which makes them not useful for improving the assembly. Thus, long read sequencing (e.g., PacBio or Nanopore) will be necessary to retrieve a more complete MSY assembly of Old World camels. Future comparisons with the alpaca MSY (Jevit et al., 2018) will be very informative to shed light on the paternal side of the evolutionary history in Old and New World camelids.

We successfully used CBacY1775 as reference to map shotgun reads from three domestic and five wild Bactrian camels and to ascertain scY SNVs (**Supplementary Table S3**). Based on 596 variants we distinguished seven haplotypes that separated into two haplogroups (W and D). Recently, Chen N. et al. (2018) investigated 29 Y-chromosomal sequence tags and 40 bovine-derived microsatellites in 94 Chinese domestic Bactrian camels. They detected the same Y-chromosomal haplotype in all tested sequence tags but one, which showed an indel, and observed allelic variation at only one microsatellite (USP9Y-STR) leading to three different HTs. This highlights the importance for further joint efforts to examine camel Y diversity.

Low MSY Sequence Variation Within Wild Bactrian Camels

The diversity on the MSY in wild two-humped camels ($\pi = 1.67^{*}10^{-6}$, without the D haplotype carrier WC305) seems to be much reduced compared to that of domestic Bactrian camels ($\pi = 1.17^{*}10^{-5}$). This may reflect the decline of the wild camel population in Mongolia over the past decades (Hare, 2008; Yadamsuren et al., 2012). The higher diversity observed in domestic compared to wild camels seems reasonable given that less than ∼2,000 wild two-humped camels are still living in Mongolia and China (Yadamsuren et al., 2012), contrary to more than 20,000 domestic Bactrian camels (FAO, 2017). Future analyses of the MSY in more wild Bactrian camels are necessary to monitor and understand the consequences of the low paternal diversity in these endangered animals, also in view of inbreeding and a successful conservation management.

Male-Driven Population Dynamics in Domestic and Wild Bactrian Camels

Based on 596 SNVs (**Supplementary Table S3**) we created a Y-phylogenetic network and observed a clear separation of wild and domestic Bactrian camels into two different haplogroups (W and D). This finding is consistent with the separation observed in mtDNA based genealogies (Silbermayr et al., 2010; Mohandesan et al., 2017). The only exception turned out to be the wild Bactrian camel WC305, which clustered with the domestic camels into haplogroup D (**Figure 2**). This male wild camel was captured during a radio-collaring mission in the GGSPAA, where officially no domestic Bactrian camels are allowed to roam. Notably, WC305 had a wild camel mtDNA haplotype (Mohandesan et al., 2017), but we found a signature of introgressed domestic alleles in this wild camel individual on the autosomal genome (Fitak et al., 2016b). So far, our MSY observation in WC305 could be interpreted a footprint of a domestic male that escaped into the wild. This inference is based on the assumption that the lineages of the populations ancestral to wild and domestic camels had differentiated before domestication. Alternatively, under the scenario of incomplete lineage sorting, the existence of wild camels carrying the D lineage would not be unexpected. The next necessary step will be haplotyping of a large number

TABLE 2 | MSY diversity estimates in domestic and wild Bactrian camels.

	θ	π
All domestic Bactrian camels	$1.17^{*}10^{-5}$	$1.17^{*}10^{-5}$
Wild Bactrian camels without WC305	$1.37^{*}10^{-6}$	$1.67^{*}10^{-6}$
All wild Bactrian camels	$1.11^{*}10^{-4}$	$1.73^{*}10^{-4}$

of male wild and domestic Bactrian camels to determine MSY haplotype frequencies. Such data should enlighten the male-driven population dynamics in these two sister species in more detail. Additional autosomal markers could also provide information on the magnitude of the hybridisation between domestic and wild Bactrian camels, which is favored by some Mongolian camel owners due to the effect of hybrid vigor (Silbermayr and Burger, 2012; Yadamsuren et al., 2012).

Divergence Time Estimates of Paternal and Maternal Genealogies in Wild and Domestic Bactrian Camels Are Not Consistent

The mitochondrial genomes of wild and domestic Bactrian camels coalesce 1.1 (0.58–1.8) mya (Mohandesan et al., 2017). This is much older than the divergence time estimate of 26,999 ± 2,268 ya based on the MSY. Inconsistencies between maternal and paternal estimates of the time to the MRCA have been described in horses (Wallner et al., 2017) and might be caused by a faster mutation rate in the mtDNA, the maintenance of older maternal lineages compared to the paternal counterpart and/or possibly be a signature of sex-specific generation times. The estimated time of the MRCA (1,764 (416 ya) of the domestic Bactrian haplogroup suggests that all MSY lineages in our dataset derive from a single ancestor that lived after the species domestication around 4,000- to 6,000 ya (Peters and von den Driesch, 1997). The very recent approximated MRCA (490 ± 353 ya) of the wild camels reflects a strong bottleneck in males following the dramatic population decline of these highly endangered animals over the last century.

Conservation of Male Lineages in Wild Bactrian Camels

Until now, no large-scale study has been performed to identify male genealogies in the last remaining and highly endangered wild two-humped camels in the Mongolian GGSPAA and in the Chinese deserts Lop Noor and Taklamakan. A comprehensive survey including a large number of male individuals will be necessary in the future to establish a more comprehensive view on the diversity, introgression patterns and divergence estimates within and between wild and domestic Bactrian camels. Here, we provide the basis for developing a SNV array for a continuous screening of paternal lineages to investigate male-driven population dynamics in Old World camels. This will contribute to conserve the integrity of the wild two-humped camel gene pool.

ETHICS STATEMENT

EDTA blood samples for domestic and wild Bactrian camels were retrieved commensally during routine veterinary controls, micro-chipping or radio-collaring of Mongolian wild camels, respectively. All data sets were collected within the frames of the legal requirements of Austria and Mongolia. Micro-chipping of semi-wild camels from the breeding center of the Wild Camel Protection Foundation (WCPF) in Mongolia was performed with the request and consent of the WCPF. Capture and collaring of wild camels within the Mongolian GGSPAA (Walzer et al., 2012) was conducted within a cooperation agreement between the International Takhi Group and the Mongolian Ministry of Nature, Environment and Tourism signed on 15.02.2001 and renewed on 27.01.2011.

AUTHOR CONTRIBUTIONS

SF performed bioinformatic analyses and lab work for variant validation. SF and BW developed the methodology and wrote and discussed the manuscript. BC, AY, BF, and CW provided samples and interpreted results. GB provided resources and discussions. PB conceived and managed the project and wrote the manuscript. All authors edited the manuscript.

FUNDING

SF was funded by the Austrian Federal Ministry of Agriculture, Forestry, Environment and Water Management (DAFNE, 101184). PB acknowledges funding from the Austrian Science Fund (FWF) P29623-B25.

ACKNOWLEDGMENTS

We are very grateful to the Wild Camel Protection Foundation for the support in sample collection. We thank the Zoo Herberstein and H. Burgsteiner, Austria for agreeing to use camel samples for research purposes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00423/full#supplementary-material

REFERENCES

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-ter Wengel, P., Pitt, D., et al. (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6707–6712. doi: 10.1073/pnas.1519508113

Bandelt, H. J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Chen, H., Ren, Z., Zhao, J., Zhang, C., and Yang, X. (2018). Y-chromosome polymorphisms of the domestic Bactrian camel in China. *J. Genet.* 97, 3–10. doi: 10.1007/s12041-017-0852-1

Chen, N., Cai, Y., Chen, Q., Li, R., Wang, K., Huang, Y., et al. (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat. Commun.* 9:2337. doi: 10.1038/s41467-018-04737-0

Elbers, J. P., Rogers, M., Perelman, P., Proskuryakova, A., Serdyukova, N., Johnson, W., et al. (2019). Improving illumina assemblies with hi-c and long reads: an example with the north african dromedary. *Mol. Ecol. Res.* [Epub ahead of print].

FAO (2017). *FAOSTAT*. Rome: Food and Agriculture Organization.

Felkel, S., Vogl, C., Rigler, D., Dobretsberger, V., Chowdhary, B. P., Distl, O., et al. (2019). The horse Y chromosome as an informative marker for tracing sire lines. *Sci. Rep.* 9:6095. doi: 10.1038/s41598-019-42640-w

Felkel, S., Vogl, C., Rigler, D., Jagannathan, V., Leeb, T., Fries, R., et al. (2018). Asian horses deepen the MSY phylogeny. *Anim. Genet.* 49, 90–93. doi: 10.1111/age.12635

Fitak, R., Mohandesan, E., Corander, J., and Burger, P. A. (2016a). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Res.* 16, 314–324. doi: 10.1111/1755-0998.12443

Fitak, R., Mohandesan, E., Corander, J., Yadamsuren, A., Chuluunbat, B., Abdelhadi, O., et al. (2016b). *Genomic Footprints of Selection Under Domestication in Old World Camelids*. Available at: https://pag.confex.com/pag/xxiv/webprogram/Paper18540.html (accessed January 09, 2016).

Hare, J. (2008). Camelus ferus. *IUCN Red List Threaten. Spec.* 2008:e.T63543A12689285.

Jevit, M., Hillhouse, A., Richardson, M., Davis, B., Juras, R., Malcolm, A., et al. (2018). "Comparative study of dromedary and alpaca Y chromosomes in recent advances in camelids biology, health and production," in *Proceedings of the 5th Conference of the International Society for Camelid Research and Development 2018*, eds A. Sghiri and F. Kichou (Rabat: Institute Agronomique et Veterinaire Hassan II), 115–116.

Jirimutu, Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., et al. (2012). Genome sequences of wild and domestic bactrian camels. *Nat. Commun.* 3:1202. doi: 10.1038/ncomms2192

Jobling, M. A., and Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* 4, 598–612. doi: 10.1038/nrg1124

Jobling, M. A., and Tyler-Smith, C. (2017). Human Y-chromosome variation in the genome-sequencing era. *Nat. Rev. Genet.* 18, 485–497. doi: 10.1038/nrg.2017.36

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488, 471–475. doi: 10.1038/nature11396

Lei, Y., Hare, J., Guoying, Y., and Yun, C. (2012). "The status of the wild camel in China," in *Camels in Asia and North Africa. Interdisciplinary Perspectives on their Significance in Past and Present*, eds E. M. Knoll and P. A. Burger (Vienna: Austrian Academy of Science Press), 55–60.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (1000). Genome project data processing subgroup. (2009). The sequence alignment/map (SAM) format and SAM tools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Mohandesan, E., Fitak, R. R., Corander, J., Yadamsuren, A., Chuluunbat, B., Abdelhadi, O., et al. (2017). Mitogenome sequencing in the genus camelus reveals evidence for purifying selection and long-term divergence between wild and domestic bactrian camels. *Sci. Rep.* 7:9970. doi: 10.1038/s41598-017-08995-8

Oetjens, M. T., Martin, A., Veeramah, K. R., and Kidd, J. M. (2018). Analysis of the canid Y-chromosome phylogeny using short-read sequencing data reveals the presence of distinct haplogroups among neolithic european dogs. *BMC Genom.* 10:350. doi: 10.1186/s12864-018-4749-z

Peters, J., and von den Driesch, A. (1997). The two-humped camel (Camelus bactrianus): new light on its distribution, management and medical treatment in the past. *J. Zool.* 242, 651–679. doi: 10.1111/j.1469-7998.1997.tb05819.x

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Silbermayr, K., and Burger, P. A. (2012). "Hybridization: a threat to the genetic distinctiveness of the last wild old world camel species," in *Camels in Asia and North Africa. Interdisciplinary Perspectives on Their Significance in Past and Present*, eds E. M. Knoll and P. A. Burger (Vienna: Austrian Academy of Science Press), 69–76.

Silbermayr, K., Orozco-terWengel, P., Charruau, P., Enkhbileg, D., Walzer, C., Vogl, C., et al. (2010). High mitochondrial differentiation levels between wild and domestic Bactrian camels: a basis for rapid detection of maternal hybridization. *Anim. Genet.* 41, 315–318. doi: 10.1111/j.1365-2052.2009.01993.x

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825–837.

Skinner, B. M., Lachani, K., Sargent, C. A., and Affara, N. A. (2013). Regions of XY homology in the pig X chromosome and the boundary of the pseudoautosomal region. *BMC Genet.* 14:3. doi: 10.1186/1471-2156-14-3

Wallner, B., Palmieri, N., Vogl, C., Rigler, D., Bozlak, E., Druml, T., et al. (2017). Y chromosome uncovers the recent oriental origin of modern stallions. *Curr. Biol.* 27, 2029–2035. doi: 10.1016/j.cub.2017.05.086

Walzer, C., Kaczensky, P., Enkhbileg, D., and Yadamsuren, A. (2012). "Working in a freezer: capturing and collaring wild bactrian camels," in *Camels in Asia and North Africa. Interdisciplinary Perspectives on Their Significance in Past and Present*, eds E. M. Knoll and P. A. Burger (Vienna: Austrian Academy of Science Press), 61–68.

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Yadamsuren, A., Dulamtseren, E., and Reading, R. P. (2012). "The conservation status and management of wild camels in Mongolia," in *Camels in Asia and North Africa. Interdisciplinary Perspectives on Their Significance in Past and Present*, eds E. M. Knoll and P. A. Burger (Vienna: Austrian Academy of Science Press), 45–54.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Felkel, Wallner, Chuluunbat, Yadamsuren, Faye, Brem, Walzer and Burger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 May 2019
doi: 10.3389/fgene.2019.00482

Comparative Analysis of the TRB Locus in the *Camelus* Genus

Rachele Antonacci[1], Mariagrazia Bellini[1], Giovanna Linguiti[1], Salvatrice Ciccarese[1] and Serafina Massari[2]*

[1] Department of Biology, University of Bari Aldo Moro, Bari, Italy, [2] Department of Biological and Environmental Sciences and Technologies, University of Salento, Lecce, Italy

T cells can be separated into two major subsets based on the heterodimer that forms their T cell receptors. αβ T cells have receptors consisting of α and β chains, while γδ T cells are composed of γ and δ chains. αβ T cells play an essential role within the adaptive immune responses against pathogens. The recent genomic characterization of the *Camelus dromedarius* T cell receptor β (TRB) locus has allowed us to infer the structure of this locus from the draft genome sequences of its wild and domestic Bactrian congeners, *Camelus ferus* and *Camelus bactrianus*. The general structural organization of the wild and domestic Bactrian TRB locus is similar to that of the dromedary, with a pool of TRBV genes positioned at the 5′ end of D-J-C clusters, followed by a single TRBV gene located at the 3′ end with an inverted transcriptional orientation. Despite the fragmented nature of the assemblies, comparative genomics reveals the existence of a perfect co-linearity between the three Old World camel TRB genomic sequences, which enables the transfer of information from one sequence to another and the filling of gaps in the genomic sequences. A virtual camelid TRB locus is hypothesized with the presence of 33 TRBV genes distributed in 26 subgroups. Likewise, in the artiodactyl species, three in-tandem D-J-C clusters, each composed of one TRBD gene, six or seven TRBJ genes, and one TRBC gene, are placed at the 3′ end of the locus. As reported in the ruminant species, a group of four functional TRY genes at the 5′ end and only one gene at the 3′ end, complete the camelid TRB locus. Although the gene content is similar, differences are observed in the TRBV functional repertoire, and genes that are functional in one species are pseudogenes in the other species. Hence, variations in the functional repertoire between dromedary, wild and domestic Bactrian camels, rather than differences in the gene content, may represent the molecular basis explaining the disparity in the TRB repertoire between the *Camelus* species. Finally, our data contribute to the knowledge about the evolutionary history of Old World camelids.

Keywords: T cell receptor, TRB locus, Old World camelids, ImMunoGeneTics database, TRY genes

OPEN ACCESS

Edited by:
Pablo Orozco-terWengel, Cardiff University, United Kingdom

Reviewed by:
Terje Raudsepp, Texas A&M University, United States
Alfredo Pauciullo, University of Turin, Italy

****Correspondence:***
Rachele Antonacci
rachele.antonacci@uniba.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *21 September 2018*
Accepted: *06 May 2019*
Published: *24 May 2019*

Citation:
Antonacci R, Bellini M, Linguiti G, Ciccarese S and Massari S (2019) Comparative Analysis of the TRB Locus in the Camelus Genus. Front. Genet. 10:482. doi: 10.3389/fgene.2019.00482

INTRODUCTION

Old World camelids consist of three extant species, including the one-humped camel or dromedary, *Camelus dromedarius*, and the two-humped wild and domestic Bactrian camels, *Camelus ferus* and *Camelus bactrianus*. All the species are adapted to live in specific weather conditions of both desert and semi-desert regions and undergo environmental pressures, potentially resulting in the

evolution of adaptations specific to each species. While the species status of the dromedary and domestic Bactrian camel has been established, the evolutionary relationship between the two-humped camels has long been debated, and only recently the wild two-humped camel has been recognized as a separate species, *Camelus ferus*, on the basis of the mitochondrial (Ji et al., 2009; Silbermayr et al., 2010; Mohandesan et al., 2017) and nuclear (Jirimutu et al., 2012; Silbermayr and Burger, 2012) genetic data.

The camelid species are an interesting model to study the immune system responsible for antigen recognition (Ciccarese et al., unpublished). Together with B cells, which produce immunoglobulin (IG), αβ and γδ T cells are the major cellular components of the adaptive immune system. A specialized genetic machinery, from simple interchanges of a small number of genes, creates great diversity in the IG and in αβ and γδ T cell receptor (TR) that potentially generate billions of different IG and TR. The adaptive immune response of camels displays characteristic features, such as heavy chain antibody homodimers in the serum (Hamers-Casterman et al., 1993; Muyldermans et al., 2009) and a limited germline repertoire of T cell receptor γ (TRG) and δ (TRD) chain genes compared to other artiodactyl species (Vaccarelli et al., 2008; Piccinni et al., 2015), diversified by the extensive somatic hypermutation (SMH) (Antonacci et al., 2011; Vaccarelli et al., 2012; Ciccarese et al., 2014). In the same way, the *Camelus dromedarius* repertoire of T cell receptor β (TRB) chain appears reduced in size and gene content compared to the other species (Antonacci et al., 2017a,b). Anyhow, the structural organization of the *Camelus dromedarius* TRB locus is similar to that of the other mammalian species, with a pool of Variable (TRBV) genes positioned at the 5′ end of Diversity (TRBD), Joining (TRBJ) and Constant (TRBC) genes, followed by a single TRBV gene, with an inverted transcriptional orientation located at the 3′ end. The TRBD, TRBJ and TRBC genes are organized in three D-J-C clusters, which is a common feature of sheep, cattle and pigs.

To broaden the understanding of the camel immune system, we characterized the structure of the TRB locus in *Camelus ferus* and *Camelus bactrianus* analyzing the draft genome sequences available in public database, using the human (Lefranc et al., 2003) and dromedary TRB locus (Antonacci et al., 2017a,b) as reference sequences.

MATERIALS AND METHODS

Genome Analyses

The *Camelus ferus* (wild Bactrian camel) TRB genomic sequence was retrieved directly from the CB1 assembly of the whole genome shotgun sequence available at GenBank (BioProject PRJNA76177). In particular, the whole TRB region of 302258 bp (gaps included) is entirely contained in the NW_006210980 unplaced genomic scaffold. The MOXD2 (monooxygenase, dopamine-beta-hydroxylase-like 2, DBH-like2) and EPHB6 (ephrin type-B receptor 6) genes, flanking, respectively, the 5′ and 3′ ends of TRB locus, were included in the analysis.

The *Camelus bactrianus* (domestic Bactrian camel) TRB genomic sequence was retrieved from the Ca_bactrianus_MBC_1.0 assembly of the whole genome shotgun sequence available at GenBank (BioProject PRJNA183605). The analyzed region comprises two principal unplaced genomic not continuous scaffolds: NW_011511605 (181382 bp) and NW_011509864 (99628 bp). MOXD2 and EPHB6, included in the analysis, are located within the NW_011511605 and the NW_011509864 scaffolds, respectively. A further BLAST search of the domestic Bactrian camel genomic assembly was performed, using, as a query, specific dromedary TRB gene sequences as well as the 3′ end and the 5′ end sequences of the NW_011511605 and NW_011509864 scaffolds, respectively. In this way, five short unplaced scaffolds were retrieved, including NW_011537499, which is 542 bp long and overlaps with the 3′ end of NW_011511605 and the 5′ end of NW_011509864, two continuous but not overlapping scaffolds, namely, NW_011541550 (877 bp) and NW_011529568 (303 bp), which contain the TRBV15 gene, NW_011514083 (2181 bp), which retains the TRBV16 gene, and NW_011511596 (685 bp) which retains the TRBV21S3 gene. The length of the whole TRB genomic sequence is 285598 bp (gaps included). All the scaffolds retrieved from the genomic assemblies are summarized in **Supplementary Table S1**.

The human and the dromedary TRB genomic sequences (Lefranc et al., 2015; Antonacci et al., 2017a,b) were used against the *Camelus ferus* and *Camelus bactrianus* genome sequences to identify the corresponding TRBV, TRBD, TRBJ, and TRBC genes based on homology, by the BLAST program.

The beginning and end of each coding exon were identified with accuracy by the presence of splice sites or the flanking recombination signal sequences (RSs) of the TRBV, TRBD and TRBJ genes. The locations of the TRB genes are provided in **Supplementary Tables S2**, **S3**. The sequence comparison also allowed for the identification and characterization of the camel trypsin-like serine protease (TRY) genes. The locations of the TRY genes are provided in **Supplementary Tables S4**, **S5**.

The PipMaker program (Schwartz et al., 2000)[1] was also used for the genomic comparative analyses of the *Camelus ferus* and *Camelus bactrianus* TRB loci with the dromedary sequence previously described (Antonacci et al., 2017a,b). Moreover, the computational analysis was conducted using the RepeatMasker for the identification of genome-wide repeats and low complexity regions[1].

Classification of the TRB Genes

Considering the percentage of nucleotide identity of the genes with respect to human and *Camelus dromedarius* and based on the genomic position within the locus, each TRB gene was classified, and the nomenclature was established according to IMGT at http://www.imgt.org/IMGTScientificChart/SequenceDescription/IMGTfunctionality.html (Lefranc, 2014; **Supplementary Tables S2**, **S3**). The TRBV genes were assigned to 26 different subgroups in *Camelus ferus* as well as in *Camelus bactrianus*, based on the percentage of nucleotide identity by using Clustal Omega alignment tool, which is available at

[1]http://www.repeatmasker.org

EMBL-EBI website[2], adopting the criterion that sequences with a nucleotide identity of more than 75% in the V-region belong to the same subgroup (Arden et al., 1995a,b). Due to the fragmented and incomplete nature of the genomic assemblies, a temporary designation was used for multigene subgroups, in which the Arabic number (for the subgroup) is followed by the letter S followed by the number of the gene in the subgroup.

The TRBD, TRBJ, and TRBC genes were annotated, according to the similarity with the D-J-C sequence of the dromedary species (Antonacci et al., 2017a,b). Each TRBJ1, TRBJ2 and TRBJ3 gene was designed by a hyphen and a number corresponding to their position in the cluster. They were all predicted to be functional except for the domestic Bactrian camel TRBJ3-5 and TRBJ2-6 genes whose sequences were incomplete (**Supplementary Tables S2**, **S3**).

Phylogenetic Analyses

The TRBV genes used for the phylogenetic analysis were retrieved from the following sequences deposited in the GEDI (for GenBank/ENA/DDBJ/IMGT/LIGM-DB) databases: NG_001333 (human TRB locus contig); NW_011591622, NW_011593440, NW_011591151, NW_011620189, NW_011616084, NW_011607149, NW_011601111, and LT837971 [dromedary TRB locus contig as characterized by (Antonacci et al., 2017a,b)]; NW_006210980 (this work) (wild Bactrian camel TRB locus contig); and NW_011511605, NW_011509864, NW_011514083 and NW_011511596 (this work) (domestic Bactrian camel TRB locus contig). We combined the nucleotide sequences of the V-REGION of the TRBV genes with the corresponding gene sequences of humans and dromedary.

The TRBC genes used for the phylogenetic analysis were retrieved from the following sequences deposited in the GEDI databases: NG_001333 (human TRB locus); AE000665 (mouse TRB locus); NW_003726086 [dog TRB locus as characterized by Mineccia et al. (2012)]; NW_003159384 [rabbit TRB locus as characterized by Antonacci et al. (2014)]; L27845 and L27844 [horse TRBC1 and TRBC2, Schrenzel et al. (1994)]; AM420900 (sheep D-J-C region, Antonacci et al., 2008); GK000004 [bovine TRB locus as characterized by Connelley et al. (2009)]; NC_010460 [pig TRB locus as characterized by Massari et al. (2018)]; LT837971 (dromedary D-J-C region, Antonacci et al., 2017b); NW_006210980 (this work) (*Camelus ferus* TRB locus); and NW_011509864 (this work) (*Camelus bactrianus* D-J-C region).

For the TRY analysis, only complete and functional genes were used. They were retrieved from the following sequences deposited in the GEDI databases: NG_001333 for three human TRY genes (PRSS58, PRSS1, and PRSS2); four dog TRY genes derived from Mineccia et al., 2012; five rabbit TRY genes derived from Antonacci et al., 2014, renamed from TRY1 to TRY5 according to the order on the TRB locus; five cattle and sheep TRY genes (TRY1-TRY5) derived from Connelley et al., 2009 and from NC_019461 (personal communication), respectively, NC_010460 for four pig genes (TRY1-TRY4 as in Massari et al., 2018); NW_011591622 and NW_011591151 for three dromedary genes (TRY1, TRY2, and TRY4) named on the basis of the bovine nomenclature; NW_006210980 for three *Camelus ferus* genes (this work); and NW_011511605 and NW_011509864 for five *Camelus bactrianus* genes (this work).

The information about genes used in the phylogenetic analyses are summarized in **Supplementary Table S6**.

Multiple alignments of the gene sequences under analysis were carried out with the MUSCLE program (Edgar, 2004). The evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). We used the neighbor-joining (NJ) method to reconstruct the phylogenetic tree (Saitou and Nei, 1987). The evolutionary distances were computed using the p-distance method (Nei and Kumar, 2000) and are in the units of the number of base differences per site.

RESULTS

Genomic Organization of the *Camelus ferus* and *Camelus bactrianus* TRB Loci as Drawn From the Assemblies

A standard BLAST search of the genomic resources was performed by using the human and dromedary TRB gene sequences to assess their location in the wild and domestic Bactrian camel genomes. We directly retrieved a sequence of 302258 bp (gaps included) from the *Camelus ferus* CB1 assembly, which corresponds to one unplaced genomic scaffold, and a sequence of 285598 bp (gaps included) from the *Camelus bactrianus* Ca_bactrianus_MBC_1.0 assembly, which corresponds to seven distinct unplaced scaffolds. In particular, NW_011511605, NW_011509864 and NW_011537499 are overlapping scaffolds and form a continuous sequence. The genes MOXD2 and EPHB6, which flank the 5′ and 3′ ends, respectively, of all the mammalian TRB loci studied to date, were comprised in the sequences. We identified and annotated all the TRB genes, taking into account, as a reference, both the human and the dromedary sequences[3], Lefranc et al., 2015; Antonacci et al., 2017a,b). The functionality of the V, D, J and C genes was predicted through the manual alignment of the sequences adopting the following parameters: (a) the identification of the leader sequence at the 5′ of the TRBV genes; (b) the determination of the proper RSs located at 3′ of the TRBV (V-RS) gene, the 5′ and 3′ ends of the TRBD (5′D-RS and 3′D-RS) and the 5′ of the TRBJ (J-RS) gene; (c) the determination of conserved acceptor and donor splicing sites; (d) the estimation of the expected length of the coding regions; and (e) the absence of frameshifts and stop codons in the coding regions of the genes.

The general structural organization of the *Camelus ferus* and *Camelus bactrianus* TRB loci follows that of other mammalian species, with a library of TRBV genes positioned at the 5′ end of the D-J-C clusters, followed by a single TRBV gene, with an inverted transcriptional orientation located at the 3′ end (**Figure 1** and **Supplementary Figure S1**).

[2]http://www.ebi.ac.uk/

[3]htpp://www.imgt.org/

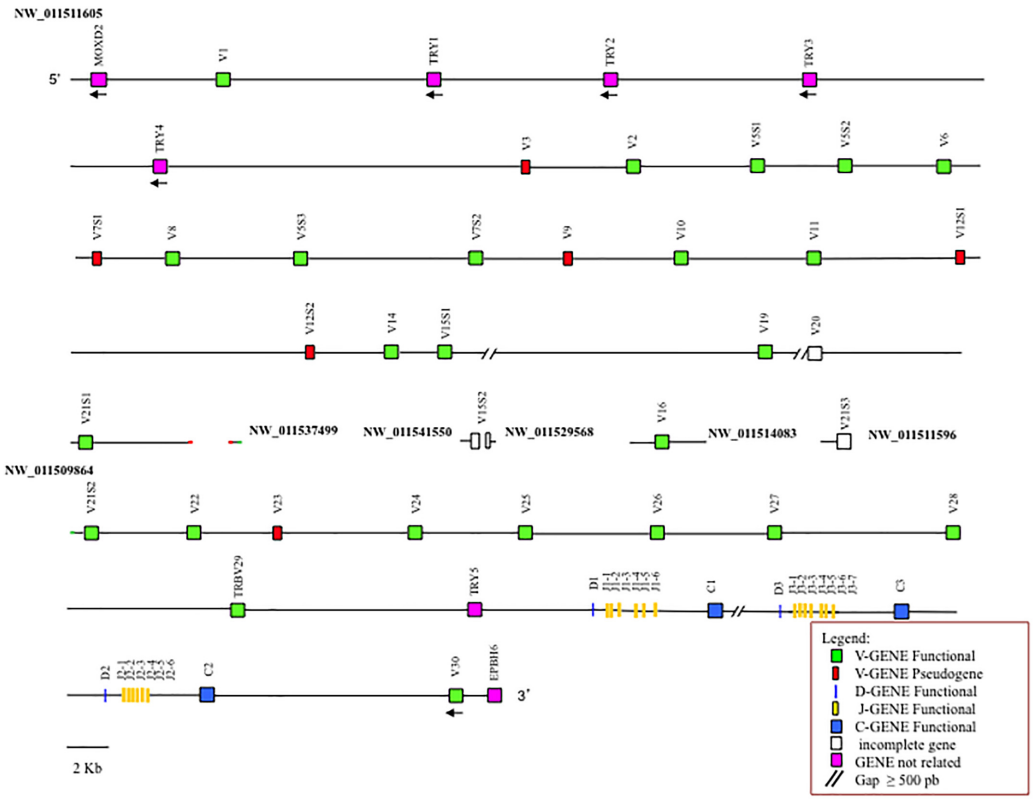


FIGURE 1 | Schematic representation of the genomic organization of the *Camelus bactrianus* TRB locus deduced from the genome assemblies. The diagram shows the position of all the related and unrelated TRB genes according to the nomenclature. The NW_011511605, NW_011537499 and NW_011509864 are overlapping scaffolds, with NW_011537499 in between NW_011511605 and NW_011509864. The boxes representing the genes are not to scale. The white boxes are incomplete genes. The exons are not shown. The arrows indicate the transcriptional orientation of the genes.

It is noteworthy that the analysis of the sequence revealed the presence of only one D-J-C cluster in *Camelus ferus* (**Supplementary Figure S1**), whereas three D-J-C clusters are present in the *Camelus bactrianus* sequence (**Figure 1**), as in *Camelus dromedarius* (Antonacci et al., 2017a,b). The fragmented and incomplete nature of the genomic assemblies might justify this discrepancy in the *Camelus ferus* genome.

In *Camelus bactrianus,* the NW_011541550, NW_011529568, NW_011514083, and NW_011511596 scaffolds can be tentatively positioned within the TRB locus, based on the *Camelus dromedarius* as well as the *Camelus ferus* sequence structures.

Classification of the *Camelus ferus* and *Camelus bactrianus* TRBV Genes and Comparison With the Dromedary Genes

We annotated 30 TRBV germline genes in the wild Bactrian camel genome, while as in dromedary (Antonacci et al., 2017a,b), 33 TRBV genes were found in the domestic Bactrian camel TRB locus. They were assigned to 26 distinct subgroups both in *Camelus ferus* and *Camelus bactrianus.*

To classify the TRBV gene subgroups, the evolutionary relationship of these genes was investigated by comparing all the wild and domestic Bactrian camel genes with the available genes corresponding to humans and dromedary by adopting two selection criteria as follows: (1) only the potential functional genes and in-frame pseudogenes (except for human TRBV1) were included, and (2) only one gene for each of the human subgroups was selected. Thus, the V-REGION nucleotide sequences of all the selected TRBV genes were combined in the same alignment, and an unrooted phylogenetic tree was made using the NJ method (Saitou and Nei, 1987; **Figure 2**). The tree shows that each of the *Camelus ferus* as well as of the *Camelus bactrianus* subgroups come together and form a monophyletic group, if present, with a corresponding human and dromedary gene. Therefore, according to phylogenetic clustering, we classified each Bactrian camel TRBV subgroup as orthologous to their corresponding human and dromedary subgroups. The only exception is the *Camelus ferus* and *Camelus bactrianus* TRBV9 genes, which were named as dromedary genes based on their genomic position within the TRB locus (Antonacci et al., 2017a,b), even if they are related to the human TRBV13 gene. Moreover, as already reported (Massari et al., 2018), the human TRBV9 is grouped together with the orthologous TRBV5S2 gene of the other mammalian species. Three human TRBV subgroups (TRBV4, TRBV17, and TRBV18) are lacking in all three camel species, indicating that these subgroups have been lost in these species or alternatively they might have originated after the separation

TABLE 1 | Correspondence between camel and human TRBV subgroup genes.

Subgroups	Human	*C. dromedarius*	*C. ferus*	*C. bactrianus*
TRBV1	1 (P)	1	1 (P)	1
TRBV2	1	1	1	1
TRBV3	2	1 (P)	1 (P)	1 (P)
TRBV4	3	–	–	–
TRBV5	8	3	3	3
TRBV6	9	1	1	1
TRBV7	9	2	1 (P) + 1	1 (P) + 1
TRBV8	2	1	1	1
TRBV9	1	(TRBV5S2)	(TRBV5S2)	(TRBV5S2)
TRBV10	3	1	1	1
TRBV11	3	1	1	1
TRBV12	5	2 (P)	1 (P)+ 1	2 (P)
TRBV13	1	1 (TRBV9 P)	1 (TRBV9 P)	1 (TRBV9 P)
TRBV14	1	1 (P)	1	1
TRBV15	1	2	1	1 + 1 (nd)
TRBV16	1	1	1 (P)	1
TRBV17	1 (ORF)	–	–	–
TRBV18	1	–	–	–
TRBV19	1	1	1	1
TRBV20	1	1	1 (ORF)	1 (nd)
TRBV21	1	2 + 1 (P)	1	2 + 1 (P)
TRBV22	1	1	1	1
TRBV23	1	1 (P)	1 (P)	1 (P)
TRBV24	1	1 (P)	1	1
TRBV25	1	1	1	1
TRBV26	1	1	1 (P)	1
TRBV27	1	1	1	1
TRBV28	1	1	1 (P)	1
TRBV29	1	1	1	1
TRBV30	1	1	1	1
TOTAL	**66**	**33**	**30**	**33**

The functionality is also reported. P, pseudogene; ORF, open reading frame; Nd, not defined (indicates that the nt sequence of the gene is incomplete and its functionality cannot be defined).

of Camelidae from the other mammalian species. Finally, all the camel TRBV1 genes group together. **Table 1** summarizes the correspondence of each TRBV gene subgroup between the camel species, with respect to the humans.

The TRBV gene functionality was defined based on the IMGT rules as described above. Twenty-one genes in *Camelus ferus* (70%) and twenty-four genes in *Camelus bactrianus* (80%) were predicted to be functional (**Supplementary Tables S2**, **S3**, **S7**). Three subgroups (TRBV5, TRBV7, and TRBV12) in the wild and four subgroups (TRBV5, TRBV7, TRBV12, and TRBV21) in the domestic Bactrian camels are multimembers, with a limited number of genes (from 2 to 3). The TRBV20 gene has an anomalous V-EXON but with an open reading frame in *Camelus ferus*, while the same gene is incomplete in *Camelus bactrianus*. The *Camelus bactrianus* TRBV21S3 gene is also incomplete. Moreover, the TRBV15S2 gene is split in the NW_011541550 and NW_011529568 scaffolds, and because of a gap in between, it lacks a portion of the V-EXON.

Nine TRBV genes in *Camelus ferus* and six in *Camelus bactrianus* are pseudogenes (**Supplementary Table S7**). Five of these (TRBV3, TRBV7S1, TRBV9, TRBV12S1, and TRBV23) are shared between the two loci. TRBV3, TRBV9, and TRBV23 are also pseudogenes in dromedary (Antonacci et al., 2017a), while the TRBV7S1 gene is functional. Moreover, the TRBV14 and TRBV24 genes are pseudogenes in dromedary, whereas they are functional in both the wild and domestic Bactrian camel. Four TRBV pseudogenes (TRBV1, TRBV16, TRBV26 and TRBV28) in *Camelus ferus* are functional in *Camelus bactrianus* and *Camelus dromedarius*. In contrast, the TRBV12S1 gene is functional in *Camelus ferus,* but is a pseudogene in *Camelus bactrianus* and *Camelus dromedarius*. We must consider that some of these discrepancies might be due to sequence errors.

The deduced amino acid sequences of the wild and domestic Bactrian camel germline TRBV genes were manually aligned together with the corresponding dromedary genes according to IMGT unique numbering for the V-REGION (Lefranc et al., 2003) to maximize the percentage of identity (**Supplementary Figure S2**). A great sequence identity between orthologous genes was observed, confirming the close relatedness of the three camel species. Mostly, the amino acid variations might be ascribed to allelic polymorphisms. Few amino acid differences (more than two residues) characterized the TRBV6, TRBV7S1, TRBV8, TRBV24 and TRBV27 genes.

Characterization of the D-J-C Region

In *Camelus ferus*, only one D-J-C cluster was detected, with one TRBD gene, seven TRBJ genes and one TRBC gene (**Supplementary Figure S1**). The nucleotide sequence comparison with the dromedary TRB genes (Antonacci et al., 2017a) revealed that the TRBD and the TRBJ genes are homologous to the corresponding genes of the D-J-C cluster 1, whereas the TRBC gene corresponds to the dromedary TRBC2 gene. Although a different organization of the D-J-C region in wild camel, with respect to the dromedary, cannot be excluded, the discrepancy is most likely due to a gap in the genomic assembly that comprises the TRBC1 gene, the entire D-J-C cluster 3, the TRBD2 and the TRBJ2 genes. This conclusion is in accordance with the presence of three D-J-C clusters (**Figure 1**) in *Camelus bactrianus* that perfectly match the structure of the *Camelus dromedarius* TRB locus except for the lack of the TRBJ3-6 gene, which is likely due to the gap present in the genomic assembly.

Moreover, the TRBJ1 cluster in *Camelus ferus* consists of seven genes, one more than in *Camelus dromedarius* as well as in *Camelus bactrianus*. The sequence analysis showed that the region of 315 bp, containing the TRBJ1-7 gene (from position 3052198 to 3052512 of NW_006210980) is identical to the TRBJ1-4 region. This redundancy indicates a probable error in the sequence assembly whereby the TRBJ1-7 gene should be excluded from the *Camelus ferus* TRB locus.

All the genes were analyzed in detail for their structure (**Supplementary Figure S3**) and functionality (**Supplementary Tables S2**, **S3**). The only exceptions are the *Camelus bactrianus* TRBJ2-6, TRB3-5 and TRBC1 genes whose functionality

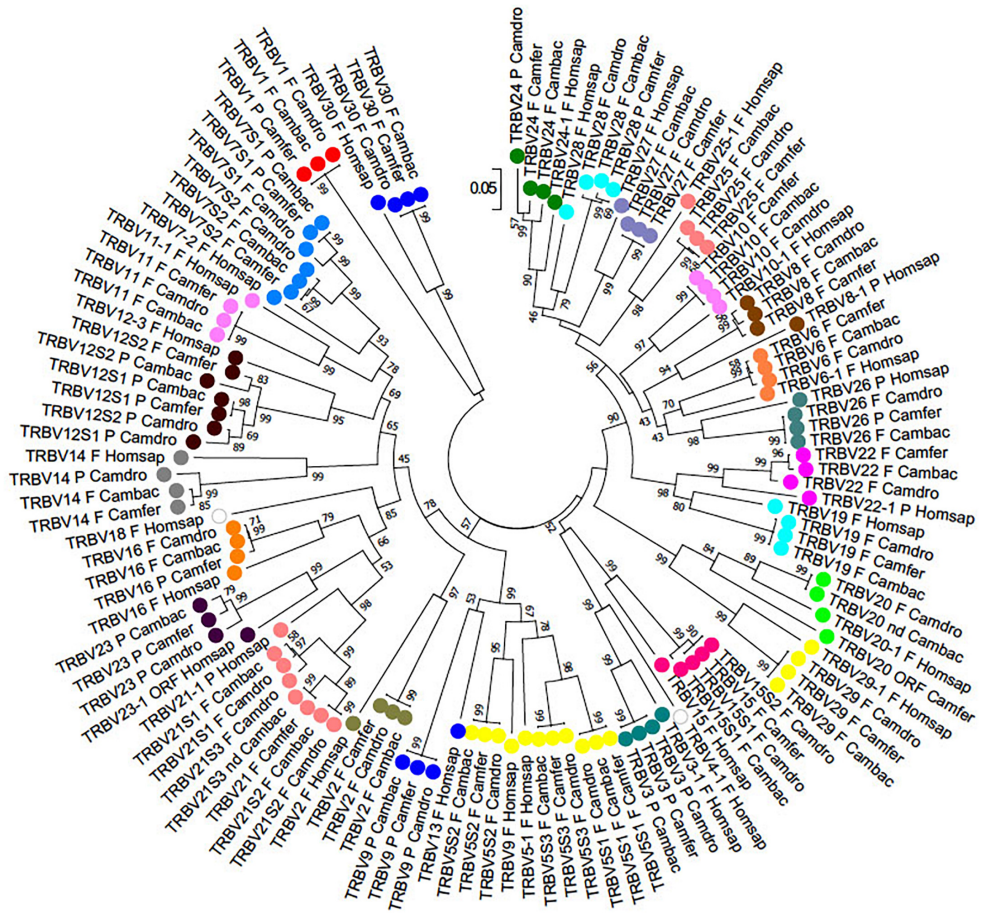


FIGURE 2 | The NJ tree inferred from the wild and domestic Bactrian camel, human and dromedary TRBV gene sequences. The evolutionary analysis was conducted in MEGA7 (Kumar et al., 2016). The optimal tree with the sum of branch length = 8.80179972 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to the branches (Felsenstein, 1985). The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer phylogenetic trees. The evolutionary distances were computed using the p-distance method (Nei and Kumar, 2000) and are in the units of the number of base differences per site. The analysis involved 123 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 130 positions in the final dataset. The different colors highlight the distribution of the phylogenetic groups. The gene functionality according to IMGT rules (F, functional, ORF, open reading frame, P, pseudogene) is indicated; "nd" indicates that the nucleotide sequence of the gene is incomplete and its functionality cannot be defined. The IMGT 6-letter for species (Homsap, Camdro, Camfer, Cambac) standardized abbreviation for taxon is used.

cannot be defined for the incomplete nucleotide sequence (**Supplementary Figures S3b,c**). The comparison between the wild and domestic Bactrian camels showed that the sequences of the orthologous genes are identical, except for the TRBJ1-6 gene, which is present in *Camelus bactrianus* with un unusual amino acid FGXG motif (FGLS), and for a polymorphism in the first exon of the TRBC2 gene (**Supplementary Figures S3b,c**).

The comparison with the dromedary genes (Antonacci et al., 2017a,b) revealed a perfect correspondence between orthologous genes; the only exception is the possible exchange between the TRBD2 and TRBD3 genes in *Camelus bactrianus,* with respect to the reference genomic sequence of *Camelus dromedarius.*

In **Supplementary Figure S3c**, the protein display of the wild and domestic Bactrian camel TRBC genes, compared with the reference sequences of the dromedary genes, is shown. All the TRBC genes encode a similar protein of 178 amino acids, with the extracellular domain encoded by exon 1, exon 2 and by the first codon of exon 3. The transmembrane region is encoded by the remaining part of exon 3, while the cytoplasmic portion is encoded by exon 4.

To gain insight into the evolution of the camel TRBC genes within Camelidae family and with respect to Artiodactyla, we combined in the same alignment the nucleotide sequences of the coding regions of the *Camelus ferus, Camelus bactrianus* and *Camelus dromedarius* TRBC genes together with those derived from cow, sheep and pig. The TRBC sequences retrieved from the human, mouse, dog, horse and rabbit TRB loci were also included in the analysis. A phylogenetic tree was constructed using the NJ method (**Figure 3**). In the tree, different from the TRBV genes, the TRBC genes are grouped in a species-specific manner, indicating a closer relationship between paralogous TRBC genes rather than orthologous genes.

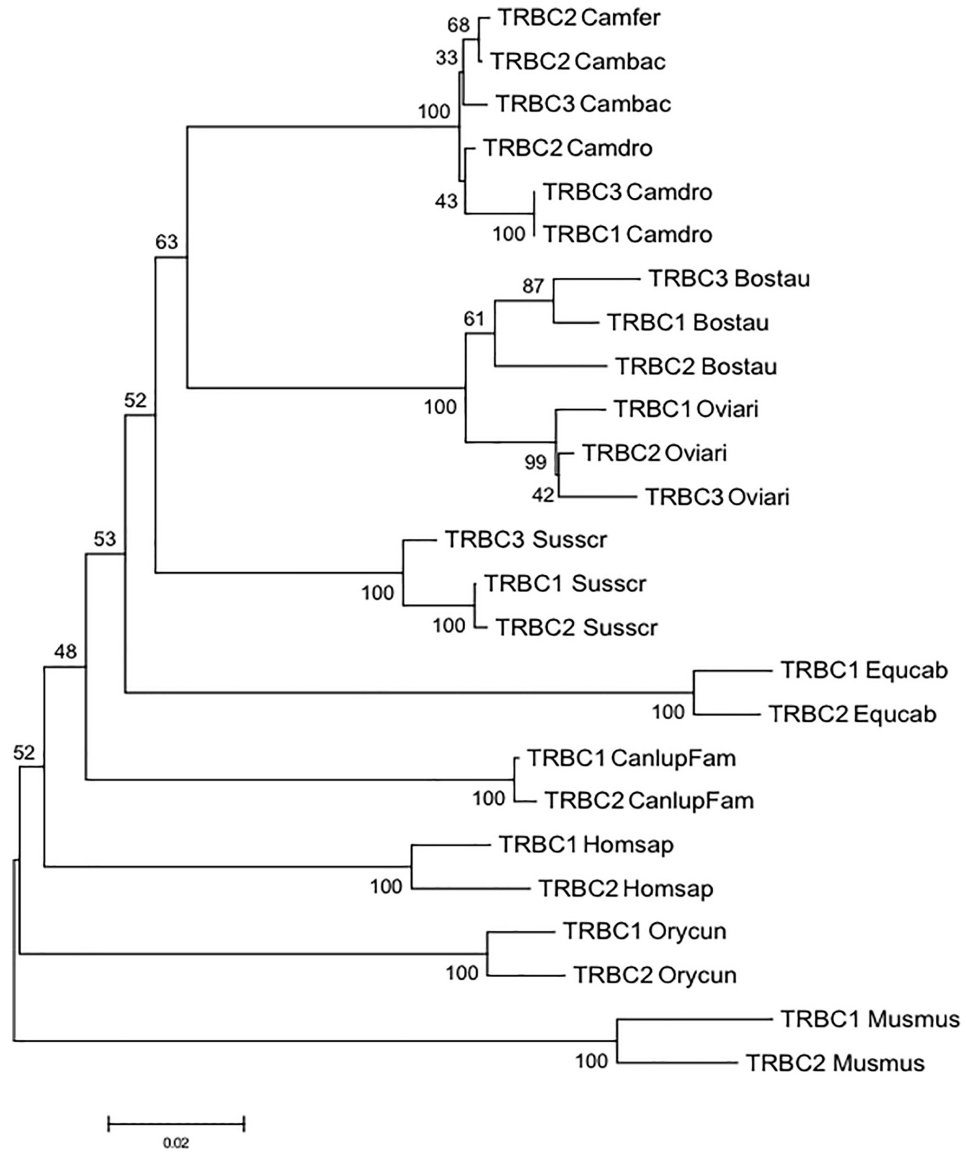


FIGURE 3 | The NJ tree inferred from the TRBC gene sequences within mammalian species. The evolutionary analysis was conducted in MEGA7 (Kumar et al., 2016). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to the branches (Felsenstein, 1985). The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer phylogenetic trees. The evolutionary distances were computed using the p-distance method (Nei and Kumar, 2000) and are in the units of the number of base differences per site. The analysis involved 25 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 515 positions in the final dataset. The IMGT 6-letter for species (Homsap, Musmus, Susscr, Bostau, Oviari, Equcab, Orycun, Camdro, Camfer, and Cambac) and 9-letter for subspecies (Canlupfam) standardized abbreviation for taxon is used.

The only exception seems to be the tight vicinity between the wild and domestic Bactrian camel TRBC2 genes. Moreover, within Artiodactyla, the camel TRBC genes form a sister group with the ruminant genes, whereas the pig genes represent a paraphyletic taxon.

Phylogenetic Analysis of the TRY Genes

The comparison of the entire wild and domestic Bactrian camel sequences to the dromedary and human sequences allowed us also to identify and annotate unrelated TRB genes, consisting of the MOXD2 and EPHB6 genes, which delimit the TRB locus, and of a group of TRY genes that are typically interspersed among mammalian TRB genes.

Downstream of the TRBV1 gene, proceeding from 5′ to 3′, we found two in *Camelus ferus* and four TRY genes in *Camelus bactrianus* (**Figure 1** and **Supplementary Figure S1**). In both genomic sequences, a further TRY gene was found before the D-J-C region.

To classify the camel TRY, the wild and domestic Bactrian gene sequences were aligned with those of human, dog, rabbit,

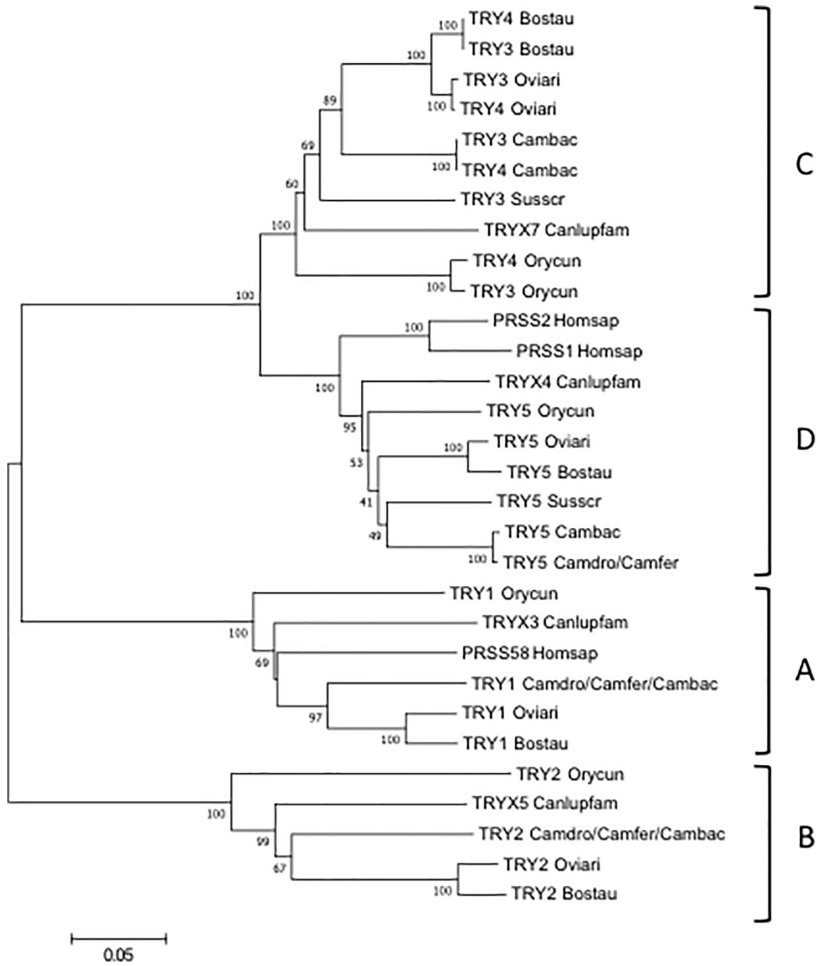


FIGURE 4 | The NJ tree inferred from the TRY gene sequences. The evolutionary analysis was conducted in MEGA7 (Kumar et al., 2016). The optimal tree with the sum of branch length = 1.55315431 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method (Nei and Kumar, 2000) and are in the units of the number of base differences per site. The analysis involved 21 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 693 positions in the final dataset. The IMGT 6-letter for species (Homsap, Susscr, Bostau, Oviari, Camdro, Camfer, and Cambac) standardized abbreviation for taxon is used.

dromedary, pig, cow and sheep and an phylogenetic tree was made (**Figure 4**).

The tree resolves the TRY genes in four monophyletic groups, each containing a *Camelus ferus* and/or *Camelus bactrianus* gene. Similar to the TRBV analysis, each group consists of orthologous TRY genes in accordance with the genomic location within the TRB locus. In A, B and C groups are present the TRY genes positioned at the 5′ region, whereas in D the TRY genes located before the D-J-C region. It is noteworthy that in C, the TRY3 and TRY4 genes of ruminants as well as of *Camelus bactrianus* are clustered in a specie-specific manner, similar to the TRBC genes, indicating their occurrence through an independent duplication event in each species or alternatively a process of sequence homogenization of the two genes within each species. A duplication event also occurred in rabbit, generating the TRY3 and TRY4 genes.

The classification, location and predicted functionality of the TRY genes are reported in **Supplementary Tables S4**, **S5**.

Genomic Comparison of the *Camelus* TRB Locus

To trace the genomic architecture of the camelid TRB locus, the TRBV region (from MOXD2 to TRY5) of the three Old World camel species were first screened with the RepeatMasker program to analyse the compositional properties (G + C content) and to identify the interspersed repeats (**Supplementary Table S8**). The GC content is approximately 45% in all three camel species. The density of the total interspersed repeats is slightly higher in *Camelus ferus* (29.68%) than *Camelus dromedarius* (28.15%) and *Camelus bactrianus* (28.82%). In all cases, the most abundant repeat elements are LINEs, whereas the percentage of SINEs

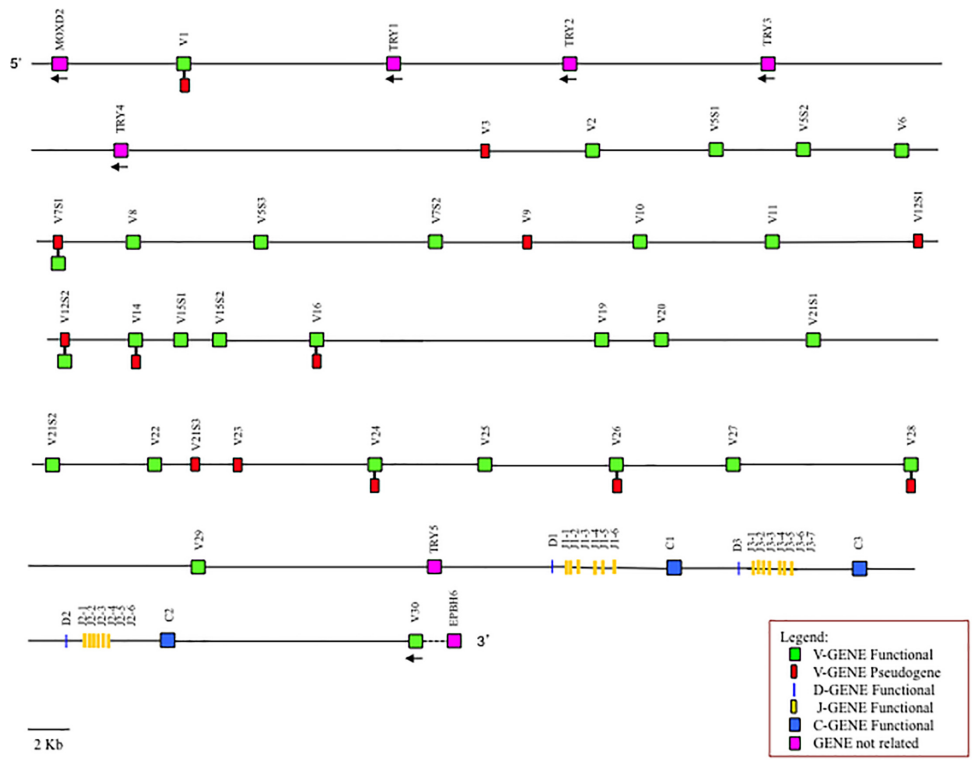


FIGURE 5 | Virtual map of the TRB locus of Old World camelids. The map was constructed according to the indications given in the text. The presence of differences in the functional status of the TRBV genes between the three species is reported.

seems to be lower than that in the human and other mammalian corresponding loci (Mineccia et al., 2012; Massari et al., 2018).

Moreover, the *Camelus ferus* and *Camelus bactrianus* sequences were aligned with the dromedary one using the PipMaker program (Schwartz et al., 2000), and the alignment is expressed as a percentage identity plot (pip) (**Supplementary Figure S4**). The presence in the pip of the superimposed lines indicates the occurrence of redundant matches along the entire region. The clearest matches correspond to all the camel TRBV genes, except for TRBV1, due to the homology among the genes. The horizontal and long continuous lines, which represent ungapped alignments, are evident along the entire region, indicating a co-linearity between the three sequences, with a high percentage of similarity. In contrast, the rare interruptions of the homology line indicate the presence of probable gaps in the corresponding genomic assembly. As an example, there is an interrupted homology line within the domestic Bactrian camel sequence in correspondence of the dromedary TRBV16 gene (**Supplementary Figure S4**), which might confirm the insertion of the NW_011514083 scaffold within the *Camelus bactrianus* TRB sequence (**Figure 1**).

DISCUSSION

The development of new sequencing methodologies has made it possible to explore many mammalian genomes, providing a significant amount of material for large-scale comparative analyses. In general, despite the presence of gaps, due to the incompleteness of genome assemblies, many genes involved in complex biological traits have been discovered and mapped, contributing to the identification of molecular basis of phenotypic differences between species.

In this perspective, the recent publication of the first complete genome sequence of species belonging to the Tylopoda suborder, i.e., dromedary (Wu et al., 2014) and its related domestic and wild Bactrian species (Jirimutu et al., 2012), is a stimulant.

In this paper, we used the recent characterization of the *Camelus dromedarius* TRB locus, encoding for the β chain of the αβ T cell receptor, to establish the genomic organization of this locus in the *Camelus ferus* and *Camelus bactrianus* genomes.

The most interesting aspect that emerges from the analysis of the TRB locus is that, unlike other mammalian species, in which the TRBV genomic region is highly variable in size, in numbers of genes and subgroups as well as in the extent of polymorphisms, this appears very similar among the Old World camelid species. The locus is flanked by the MOXD2 gene at 5′ end and the EPHB6 gene at 3′ end. It consists of a pool of homologous TRBV genes, 30 genes in *Camelus ferus* and 33 in *Camelus bactrianus* as in *Camelus dromedarius*, which, in all cases, were assigned to 26 different subgroups distributed in approximately 240 Kb (**Figure 1** and **Supplementary Figure S1**; Antonacci et al., 2017a,b).

Three TRY genes in the wild Bactrian camel, as well as in dromedary, and five genes in the domestic Bactrian camel complete the region.

The phylogenetic analysis of the TRBV genes revealed an inter-species clustering of the orthologous genes with a closer relationship between the camelid species (**Figure 2**). This allowed for a classification of the TRBV genes in *Camelus ferus* and *Camelus bactrianus*, with respect to *Camelus dromedarius* as well as to humans. Overall, in all the three camelid TRB loci there are three TRBV subgroups (TRBV4, TRBV17 and TRBV18) that are missing compared to human, and only five subgroups (TRBV5, TRBV7, TRBV12, TRBV15 and TRBV21 subgroups) are multimembers, with a limited number of genes (from 2 to 3) (**Table 1**). Therefore, it is not the number of the TRBV subgroups but rather the absence of extensive duplications within the individual subgroups to make the camelid TRB locus the one with the least number of TRBV germline genes compared to those of the other mammalian species studied thus far (Connelley et al., 2009; Mineccia et al., 2012; Antonacci et al., 2014; Massari et al., 2018). A limited germline TRV number has already been described in the dromedary TRG locus, although a somatic hypermutation mechanism increases the repertoire diversity of the γ and δ chains (Antonacci et al., 2011; Vaccarelli et al., 2012; Ciccarese et al., 2014).

The classification of the TRBV pseudogenes is more articulated, since substantial differences can be observed between the camelid species (**Table 1** and **Supplementary Table S7**). Apart from the TRBV genes classified as pseudogenes in all the three species (TRBV3, TRBV9, TRBV12S1, TRBV21S3 and TRBV23), there are TRBV genes classified as pseudogenes in one species and functional in the other two (TRBV14, TRBV16, TRBV24, TRBV26 and TRBV28). Moreover, in two cases, TRBV7S1 and TRBV12S1, they are pseudogenes in two species but functional in the other. Hence, variations in the functional repertoire, rather than differences in the gene content, represent the molecular basis for the disparity in the TRBV germline repertoire between the Old World camelid species.

Likewise, as with the TRBV genes, the TRY genes are also phylogenetically grouped in an inter-species manner, with the orthologous ones of the other species (**Figure 4**). Interestingly, a duplication occurred in ruminants (cattle and sheep) and in *Camelus bactrianus* as well as in rabbit (Antonacci et al., 2014), which originated the two very similar TRY3 and TRY4 genes located at the 5′ end of the TRB locus in this species, suggesting that this organization might be shared by herbivorous mammal species. In contrast, the pig and dog TRB loci have a unique corresponding TRY gene, TRY3 and TRYX7, at the 5′ end of the TRB locus (Massari et al., 2018; Mineccia et al., 2012).

When we compare the entire TRBV genomic sequences of the Old World camelids, we noted a perfect co-linearity among them, in spite of the fragmented nature of the assemblies, which is distributed throughout the TRB genes as well as the intragenic regions (**Supplementary Figure S4**). We exploited this co-linearity to speculate on the genomic organization of the TRB locus in the *Camelus* genus that is able to overcome the gaps present in the individual camelid genomes. Putting together all the collected data, we proposed a virtual map of the TRB locus (**Figure 5**), adopting the criterion that if a situation is shared by all or at least by two out of the three camelid species it can be considered reliable. Hence, the Old World camelid TRB locus extends over a region of approximately 300 kb. Realistically, 33 TRBV genes, belonging to 26 subgroups, are placed before three in-tandem D-J-C clusters, each composed of one TRBD gene, six or seven TRBJ genes, and one TRBC gene, followed by a single TRBV gene, with an inverted transcriptional orientation located at the 3′ end. According to the tight evolutionary relationship with ruminant genes (**Figure 4**), five functional TRY genes are presumably interspersed among the TRBV genes, four are situated after the first TRBV gene and one is situated before the D-J-C cluster 1. Differences in the functional aspects of the TRBV genes among the three species were considered and are reported in the map, since they might represent allelic polymorphisms.

CONCLUSION

In conclusion, our study highlights the presence of a limited germline TRBV repertoire in *Camelus ferus* and *Camelus bactrianus* as in *Camelus dromedarius* (Antonacci et al., 2017a,b), even though there are wide and diversified TRBD and TRBJ repertoires due the presence of three D-J-C clusters. This is the substantial difference with the closest relatives, i.e., cattle (Connelley et al., 2009) and pig (Massari et al., 2018), where a consistent number of TRBV genes, which occurred by duplications, with a high degree of heterozygosis, precede the three D-J-C clusters.

Altogether, our data, improved the information on the genetic background, allowing to acquire knowledge on the evolutionary history of the Old World camelids.

AUTHOR CONTRIBUTIONS

RA, SC, and SM designed research and wrote the manuscript. RA, SM, MB, and GL contributed to the genomic analysis. All authors have read and approved the final manuscript.

FUNDING

The financial support of the University of Bari and of University of Salento is gratefully acknowledged.

ACKNOWLEDGMENTS

We thank Angela Pala for technical assistance in the manuscript preparation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00482/full#supplementary-material

FIGURE S1 | Schematic representation of the genomic organization of the *Camelus ferus* TRB locus deduced from the genome assemblies. The diagram shows the position of all the related and unrelated TRB genes according to nomenclature. The boxes representing the genes are not to scale. The exons are not shown. The arrows indicate the transcriptional orientation of the genes.

FIGURE S2 | The IMGT Protein display of the wild and domestic Bactrian camel together with the dromedary TRBV genes. Only functional genes and in-frame pseudogenes are shown. The description of the strands and loops and of the FR-IMGT and CDR-IMGT is according to the IMGT unique numbering for V-REGION (Lefranc et al., 2003). The five conserved amino acids of the V-DOMAIN (1st-CYS 23, except for the TRBV6 pseudogene, CONSERVED-TRP 41, hydrophobic AA 89, 2nd-CYS 104 and J-PHE 118) are indicated in bold. The amino acid length of the CDR-IMGT is also indicated in square brackets.

FIGURE S3 | Nucleotide and deduced amino acid sequences of the *Camelus ferus* and *Camelus bactrianus* TRBD **(a)**, TRBJ **(b)** and TRDC **(c)** genes. The consensus sequence of the heptamer and nonamer is provided at the top of the figure and is underlined. The numbering adopted for the gene classification is reported on the left of each gene. In **(a)**, the nucleotide and deduced amino acid sequences of the TRBD genes in the three coding frames are reported. They consist of a 12 bp (TRBD1), 20 bp (TRBD3) and 15 bp (TRBD2) G-rich stretches. The TRBD1 and TRBD2 genes can be productively read through their three coding phases, whereas the TRBD3 gene can be read only in two phases. The RSs that flank the 5′ and 3′ sides of the coding region are well conserved with respect to the consensus. In **(b)**, the nucleotides and deduced amino acid sequences of all the TRBJ genes are reported. The donor splice site for each TRBJ is shown. The canonical FGXG amino acid motifs, whose presence defines the functional J genes, are underlined. The unusual TRBJ1-6 gene motif is in bold. The TRBJ1-7 redundant gene is in italics. In **(c)**, the IMGT Protein display of the wild, domestic Bactrian and dromedary camel TRBC genes. The descriptions of the strands and loops were collected according to the IMGT unique numbering for the C-DOMAIN (Lefranc et al., 2005). The amino acid differences among the orthologous genes are shaded in gray.

FIGURE S4 | Pip of the TRBV region between dromedary, wild and domestic Bactrian camels. The dromedary sequence is shown on the horizontal axis, and the percentage identity to the wild (Camfer) and domestic (Cambac) Bactrian camels are shown on the vertical axis. The position and orientation of all the genes and repetitive sequences are indicated. The horizontal lines represent the ungapped alignments at the percentage similarity corresponding to the scale on the right.

TABLE S1 | Scaffolds of the camel genomic assemblies used in the analysis.

TABLE S2 | Description of the TRB genes in the *Camelus ferus* genome assembly. The position of all genes and their classification and functionality are reported.

TABLE S3 | Description of the TRB genes in the *Camelus bactrianus* genome assembly. The position of all genes and their classification and functionality are reported.

TABLE S4 | Description of the unrelated TRB genes in the *Camelus ferus* genome. The position of all genes and their classification are also reported.

TABLE S5 | Description of the unrelated TRB genes in the *Camelus bactrianus* genome. The position of all genes and their classification are also reported.

TABLE S6 | GEDI ID of the genes used in the phylogenetic analysis.

TABLE S7 | Description of *Camfer* and *Cambac* TRBV ORF and pseudogenes

TABLE S8 | Output file of the program RepeatMasker. Summary of the repeat content of *Camelus dromedarius*, *Camelus ferus*, and *Camelus bactrianus* TRBV region is shown. The GC levels are also indicated.

REFERENCES

Antonacci, R., Bellini, M., Castelli, V., Ciccarese, S., and Massari, S. (2017a). Data charactering the genomic structure of the T cell receptor (TRB) locus in *Camelus dromedarius*. *Data Brief* 14, 507–514. doi: 10.1016/j.dib.2017.08.002

Antonacci, R., Bellini, M., Pala, A., Mineccia, M., Hassanane, M. S., Ciccarese, S., et al. (2017b). The occurrence of three D-J-C clusters within the dromedary TRB locus highlights a shared evolution in tylopoda, ruminantia and suina. *Dev. Comp. Immunol.* 76, 105–119. doi: 10.1016/j.dci.2017.05.021

Antonacci, R., Di Tommaso, S., Lanave, C., Cribiu, E. P., Ciccarese, S., and Massari, S. (2008). Organization, structure and evolution of 41 kb of genomic DNA spanning the D-J-C region of the sheep TRB locus. *Mol. Immunol.* 45, 493–509. doi: 10.1016/j.molimm.2007.05.023

Antonacci, R., Giannico, F., Ciccarese, S., and Massari, S. (2014). Genomic characteristics of the T cell receptor (TRB) locus in the rabbit (*Oryctolagus cuniculus*) revealed by comparative and phylogenetic analyses. *Immunogenetics* 66, 255–266. doi: 10.1007/s00251-013-0754-1

Antonacci, R., Mineccia, M., Lefranc, M. P., Ashmaoui, H. M. E., Lanave, C., Piccinni, B., et al. (2011). Expression and genomic analyses of *Camelus dromedarius* T cell receptor delta (TRD) genes reveal a variable domain repertoire enlargement due to CDR3 diversification and somatic mutation. *Mol. Immunol.* 48, 1384–1396. doi: 10.1016/j.molimm.2011.03.011

Arden, B., Clark, S. P., Kabelitz, D., and Mak, T. W. (1995a). Human T-cell receptor variable gene segment families. *Immunogenetics* 42, 455–500.

Arden, B., Clark, S. P., Kabelitz, D., and Mak, T. W. (1995b). Mouse T-cell receptor variable gene segment families. *Immunogenetics* 42, 501–530.

Ciccarese, S., Vaccarelli, G., Lefranc, M. P., Tasco, G., Consiglio, A., Casadio, R., et al. (2014). Characteristics of the somatic hypermutation in the *Camelus dromedarius* T cell receptor gamma (TRG) and delta (TRD) variable domains. *Dev. Comp. Immunol.* 46, 300–313. doi: 10.1016/j.dci.2014.05.001

Connelley, T., Aerts, J., Law, A., and Morrison, W. I. (2009). Genomic analysis reveals extensive gene duplication within the bovine TRB locus. *BMC Genomics* 10:192. doi: 10.1186/1471-2164-10-192

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* 5:113. doi: 10.1186/1471-2105-5-113

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hamers, C., Songa, E. B., et al. (1993). Naturally-occurring antibodies devoid of light-chains. *Nature* 363, 446–448. doi: 10.1038/363446a0

Ji, R., Cui, P., Ding, F., Geng, J., Gao, H., Zhang, H., et al. (2009). Monophyletic origin of domestic bactrian camel (*Camelus bactrianus*) and its evolutionary relationship with the extant wild camel (*Camelus bactrianus ferus*). *Anim. Genet.* 40, 377–382. doi: 10.1111/j.1365-2052.2008.01848.x

Jirimutu, E., Wang, Z., Ding, G. H., Chen, G. L., Sun, Y. M., Sun, Z. H., et al. (2012). Genome sequences of wild and domestic bactrian camels. *Nat. Commun.* 3:1202. doi: 10.1038/ncomms2192

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lefranc, M. P. (2014). Immunoglobulin and T cell receptor genes: IMGT (R) and the birth and rise of immunoinformatics. *Front. Immunol.* 5:22. doi: 10.3389/fimmu.2014.00022

Lefranc, M. P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., et al. (2015). IMGT®, the international imMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* 43, D413–D422.

Lefranc, M. P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., et al. (2005). IMGT unique numbering for immunoglobulin and T cell receptor

constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.* 29, 185–203. doi: 10.1016/j.dci.2004.07.003

Lefranc, M. P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., et al. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* 27, 55–77. doi: 10.1016/s0145-305x(02)00039-3

Massari, S., Bellini, M., Ciccarese, S., and Antonacci, R. (2018). Overview of the Germline and Expressed Repertoires of the TRB Genes in *Sus scrofa*. *Front. Immunol.* 9:2526. doi: 10.3389/fimmu.2018.02526

Mineccia, M., Massari, S., Linguiti, G., Ceci, L., Ciccarese, S., and Antonacci, R. (2012). New insight into the genomic structure of dog T cell receptor beta (TRB) locus inferred from expression analysis. *Dev. Comp. Immunol.* 37, 279–293. doi: 10.1016/j.dci.2012.03.010

Mohandesan, E., Fitak, R. R., Corander, J., Yadamsuren, A., Chuluunbat, B., Abdelhadi, O., et al. (2017). Mitogenome sequencing in the genus camelus reveals evidence for purifying selection and long-term divergence between wild and domestic bactrian camels. *Sci. Rep.* 7:9970. doi: 10.1038/s41598-017-08995-8

Muyldermans, S., Baral, T. N., Retarnozzo, V. C., De Baetselier, P., De Genst, E., Kinne, J., et al. (2009). Camelid immunoglobulins and nanobody technology. *Vet. Immunol. Immunopathol.* 128, 178–183. doi: 10.1016/j.vetimm.2008.10.299

Nei, M., and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Piccinni, B., Massari, S., Jambrenghi, A. C., Ginnico, F., Lefranc, M. P., Ciccarese, S., et al. (2015). Sheep (Ovis aries) T cell receptor alpha (TRA) and delta (TRD) genes and genomic organization of the TRA/TRD locus. *BMC Genomics* 16:709. doi: 10.1186/s12864-015-1790-z

Saitou, N., and Nei, M. (1987). The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Schrenzel, M. D., Watson, J. L., and Ferrick, D. A. (1994). Characterization of horse (*Equus caballus*) T-cell receptor beta chain genes. *Immunogenetics* 40, 135–144.

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., et al. (2000). PipMaker - a web server for aligning two genomic DNA sequences. *Genome Res.* 10, 577–586. doi: 10.1101/gr.10.4.577

Silbermayr, K., and Burger, P. A. (2012). "Hybridization:a treat to the Genetic Distinctiveness of the Last Wild old World Camel Species," in *Camels in Asia and North Africa. Interdisciplinary Perspectives on Their Significance in Past and Present*, eds E. M. Knoll and P. A. Burger (Vienna: Austrian Academy of Sciences), 69–76.

Silbermayr, K., Orozco-terWengel, P., Charruau, P., Enkhbileg, D., Walzer, C., Vogl, C., et al. (2010). High mitochondrial differentiation levels between wild and domestic bactrian camels: a basis for rapid detection of maternal hybridization. *Anim. Genet.* 41, 315–318. doi: 10.1111/j.1365-2052.2009.01993.x

Vaccarelli, G., Antonacci, R., Tasco, G., Yang, F. T., Giordano, L., El Ashmaoui, H. M., et al. (2012). Generation of diversity by somatic mutation in the *Camelus dromedarius* T-cell receptor gamma variable domains. *Eur. J. Immunol.* 42, 3416–3428. doi: 10.1002/eji.201142176

Vaccarelli, G., Miccoli, M. C., Antonacci, R., Pesole, G., and Ciccarese, S. (2008). Genomic organization and recombinational unit duplication-driven evolution of ovine and bovine T cell receptor gamma loci. *BMC Genomics* 9:81. doi: 10.1186/1471-2164-9-81

Wu, H. G., Guang, X. M., Al-Fageeh, M. B., Cao, J. W., Pan, S. K., Zhou, H. M., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Antonacci, Bellini, Linguiti, Ciccarese and Massari. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fgene.2019.00445

Phylogeography and Population Genetics of *Vicugna vicugna*: Evolution in the Arid Andean High Plateau

Benito A. González[1,2], Juan P. Vásquez[3,4], Daniel Gómez-Uchida[4,5], Jorge Cortés[3,4], Romina Rivera[3,6], Nicolas Aravena[3], Ana M. Chero[3], Ana M. Agapito[3], Valeria Varas[7], Jane C. Wheleer[2,8], Pablo Orozco-terWengel[9] and Juan Carlos Marín[3]**

[1] Laboratorio de Ecología de Vida Silvestre, Facultad de Ciencias Forestales y de la Conservación de la Naturaleza, Universidad de Chile, Santiago, Chile, [2] South American Camelid Specialist Group, Survival Species Commission, International Union for Conservation of Nature, Santiago, Chile, [3] Laboratorio de Genómica y Biodiversidad, Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad del Bío-Bío, Chillán, Chile, [4] GEECLAB, Departamento de Zoología, Facultad de Ciencias Naturales y Oceanográficas, Universidad de Concepción, Concepción, Chile, [5] Núcleo Milenio INVASAL, Concepción, Chile, [6] Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad Santo Tomás, Iquique, Chile, [7] Doctorado en Ciencias, Mencioìn Ecologiìa y Evolucioìn, Instituto de Ciencias Ambientales and Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile, [8] CONOPA-Instituto de Investigación y Desarrollo de Camélidos Sudamericanos, Lima, Peru, [9] School of Biosciences, College of Biomedical and Life Sciences, Cardiff University, Cardiff, United Kingdom

OPEN ACCESS

Edited by:
Edward Hollox, University of Leicester, United Kingdom

Reviewed by:
Pierpaolo Maisano Delser, Trinity College Dublin, Ireland
Elie Poulin, Universidad de Chile, Chile

***Correspondence:**
Pablo Orozco-terWengel
Orozco-terWengelPA@cardiff.ac.uk
Juan Carlos Marín
jcmarin@biobio.cl

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *26 November 2018*
Accepted: *29 April 2019*
Published: *06 June 2019*

Citation:
González BA, Vásquez JP, Gómez-Uchida D, Cortés J, Rivera R, Aravena N, Chero AM, Agapito AM, Varas V, Wheleer JC, Orozco-terWengel P and Marín JC (2019) Phylogeography and Population Genetics of Vicugna vicugna: Evolution in the Arid Andean High Plateau. Front. Genet. 10:445. doi: 10.3389/fgene.2019.00445

The vicuña (*Vicugna vicugna*) is the most representative wild ungulate of the high Andes of South America with two recognized morphological subspecies, *V. v. mensalis* in the north and *V. v. vicugna* in the south of its distribution. Current vicuña population size (460,000–520,000 animals) is the result of population recovery programs established in response to 500 years of overexploitation. Despite the vicuña's ecosystemic, economic and social importance, studies about their genetic variation and history are limited and geographically restricted. Here, we present a comprehensive assessment of the genetic diversity of vicuña based on samples collected throughout its distribution range corresponding to eleven localities in Peru and five in Chile representing *V. v. mensalis*, plus four localities each in Argentina and Chile representing *V. v. vicugna.* Analysis of mitochondrial DNA and microsatellite markers show contrasting results regarding differentiation between the two vicuña types with mitochondrial haplotypes supporting subspecies differentiation, albeit with only a few mutational steps separating the two subspecies. In contrast, microsatellite markers show that vicuña genetic variation is best explained as an isolation by distance pattern where populations on opposite ends of the distribution present different allelic compositions, but the intermediate populations present a variety of alleles shared by both extreme forms. Demographic characterization of the species evidenced a simultaneous and strong reduction in the effective population size in all localities supporting the existence of a unique, large ancestral population (effective size ∼50,000 individuals) as recently as the mid-Holocene. Furthermore, the genetic variation observed across all localities is better explained by a model of gene flow

interconnecting them rather than only by genetic drift. Consequently, we propose space "continuous" Management Units for vicuña as populations exhibit differentiation by distance and spatial autocorrelation linked to sex biased dispersal instead of population fragmentation or geographical barriers across the distribution.

Keywords: camelids, vicuña, d-loop, microsatellites, subspecies

INTRODUCTION

The vicuña (*Vicugna vicugna*) is the most representative wild ungulate of the Andean high plateau in South America (Franklin, 1983; Wheeler, 1995, 2012). Its current distribution is limited to extreme altitude environments, living in arid landscapes with intense solar radiation and a hypoxic atmosphere (Franklin, 1982). Two morphological subspecies have been described, with geographical and habitat differences and supported by mitochondrial DNA markers (Marín et al., 2007a,b; Casey et al., 2018) the northern vicuña (*V. v. mensalis*) and the southern vicuña (*V. v. vicugna*). The subspecies *mensalis*, inhabits the 'moist puna,' is smaller and darker than the southern vicuña and is distinguished primarily by the long growth of hair on the chest (Wheeler, 2012). In contrast, the subspecies *vicugna* inhabits 'dry puna' within the Dry Diagonal belt (24° and 29° S; Ammann et al., 2001; Kull et al., 2002), lacks the long chest hairs, and has a lighter beige pelage coloration with white covering a greater portion of the body, rising halfway up the sides to mid-rib height and all the way to the ileum crest (Wheeler, 2012). Finally, the greater length of the southern vicuña molar line supports phenotypic differentiation (Wheeler and Laker, 2009). This division into two groups is further supported by the presence of two mitochondrial lineages differentiating each subspecies (Marín et al., 2017), with the southern subspecies showing greater haplotypic diversity than the northern one (Marín et al., 2007b). The biogeographical barrier between the subspecies has been suggested to correspond to the deep valley of Tarapaca in Chile (19°S) on the western side of the vicuña distribution, however, there is no evident barrier at a similar latitude on the eastern side (Bolivia) of the species distribution range (Wallace et al., 2010). Current vicuña distribution covers an area of 300,000 km^2 with several populations having increased their numbers after a drastic historic reduction (Acebes et al., 2018). Distribution is limited to altitudes from ∼3,000 to ∼5,000 m above sea level (Baigún et al., 2008; Villalba et al., 2010) along a 2,600 km stretch of the Central High Andes between 9° 30′ S in Ancash, Peru, and 27° 31′ S in the San Guillermo Reserve, Argentina (Wheeler and Laker, 2009). During the 1980s Chile, Peru, and Bolivia donated vicuña to Ecuador which were introduced to Chimborazo National Park (1° 31′ S, 78° 51′ W) and currently represent a stable, growing population (Rodriìguez and Morales-Delanuez, 2017; Vicuña Convention, 2017). This population should not be considered part of the natural vicuña distribution as there is no sound evidence that the species previously existed in Ecuador.

Current vicuña distribution and abundance is the result of population recovery programs established in response to 500 years of overexploitation (Yacobaccio, 2009) and near extinction in the 1960s (Wheeler and Laker, 2009). At the time of lowest population size only 6,000–10,000 vicuñas were left, widely distributed in low density, highly dispersed populations, with some small groups persisting at the species' southern distribution range (Grimwood, 1969; Boswall, 1972; Jungius, 1972). Thanks to the establishment of national parks and reserves, the Andean Vicuña Convention agreement, and funds from international NGOs, the vicuña population notably increased to over 200,000 individuals in four decades (Wheeler, 2006), with the northern populations showing greater recovery than the southern ones (Wheeler and Laker, 2009). Currently, a total population about 460,000–520,000 individuals inhabit the Andean high plateau (Vicuña Convention, 2017; Acebes et al., 2018), corresponding to a 50-fold increase in five decades of intensive protection and management.

Studies of vicuña genetics are limited and geographically restricted. Genetic structure has been determined using both nuclear and mitochondrial DNA in Peruvian localities (Wheeler et al., 2001), the north of Chile and Bolivia (Sarno et al., 2004), and north-western Argentina (Anello et al., 2016). The results of these studies have been used to identify four discrete Management Units (MUs; Wheeler et al., 2001; Casey et al., 2018) for the maintenance of locally adapted populations in Peru. MUs are defined as demographically independent populations whose dynamics depend on local birth and death rates rather than immigration from other populations (Taylor and Dizon, 1999). Although several researchers have advocated the use of MUs (e.g., Mosa, 1987; Marín et al., 2013b; Moodley et al., 2017), this approach has not been used for the vicuña despite its ecological, cultural and conservation importance (Wheeler et al., 2001; Sarno et al., 2004). Practical aspects deriving from the implementation of such classification would facilitate determination of the origin of skin and fiber from confiscated illegally hunted and trafficked materials (González et al., 2016), as has been done for other species (Moodley et al., 2017). Additionally, a thorough characterization of the vicuña genetic variation would facilitate comparison between the genetic diversity of wild populations and managed, captive production groups (Stølen et al., 2009; Escalante et al., 2014; Anello et al., 2016).

Here, we present a comprehensive assessment of the molecular diversity of vicuña based on samples collected throughout its distribution range. We analyze their genetic variation using 15 microsatellite loci and sequences of the left domain of the mitochondrial control region. We present evidence of (1) range-wide phylogeographic structure linked with vicuña evolutionary history; (2) patterns of molecular genetic structure

among vicuñas; and (3) links between patterns of genetic variation with phylogeographic history and barriers to gene flow. We further utilize this evidence to (4) describe and contrast the evolutionary history and patterns of gene flow among these populations in order to propose effective MUs for the species at broad scale.

MATERIALS AND METHODS

Ethics Statement

Samples were collected throughout the current distributional range of the vicuña (**Table 1** and **Figure 1**) following guidelines of the American Society of Mammalogists (Sikes et al., 2011). Specific permits were required for the Servicio Agrícola y Ganadero, SAG (permit 447, 2002), the Corporación Nacional Forestal, CONAF (permit 6/02, 2002), for granting other collection permits and helping in collecting samples. The animal research oversight committee of Universidad del Bío-Bío had knowledge of sampling plans prior to their approval of the present animal research protocol. All experimental protocols were approved by the Institutional Animal Care and Use Committee of Universidad del Bío-Bío, the methods were carried out in accordance with the approved guidelines. Samples were collected and exported for analysis (CITES permits 6282, 4222, 6007, 5971, 0005177, 0005178, 023355, 022967, and 022920) and imported to the United Kingdom (permits 269602/01, 262547/02). Peruvian samples were collected under permits from CONACS (28 September 1994, 15 June 1997), INRENA (011-c/c-2004-INRENA-IANP; 012-c/c-2004-INRENA-IANP; 016-c/c-2004-INRENA-IFFS-DCB; 016-c/c-2004-INRENA-IFFS-DCB; 021-c/c-2004-INRENA-IFFS-DCB; 026-c/c-2005-INRENA-IANP) and DGFFS (109-2009-AG-DGFFS-DGEFFS).

Sample Collection and DNA Extraction

Three hundred and fifty-three samples were collected between 1994 and 2004 at eleven Peruvian and five Chilean localities currently designated as *V. v. mensalis;* as well as four Argentine and four Chilean localities currently designated as *V. v. vicugna* (**Figure 1** and **Table 1**). Samples comprised skin (n = 4), muscle (n = 2), blood (n = 333), and feces (n = 37). All samples were stored at −80°C in the Laboratorio de Genómica y Biodiversidad, Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad del Bío-Bío, Chillán, Chile or at CONOPA in Lima, Peru. Total genomic DNA was extracted from blood using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI,

TABLE 1 | Summary of the *Vicugna vicugna* samples, including localities (ordered north to south), sample type (B, blood; F, fecal; M, muscle; S, skin), number of samples analyzed from each locality for each genetic marker.

Region	Localities; Country (abbreviation)	Sample type	Samples mtDNA (*N* = 353)	Samples microsatellites (*N* = 307)
Northern region		–	241	179
	Catac, Ancash; Perú (**CT**)	B	14	11
	Tinco Cancha, Junín; Perú (**TC**)	B	16	6
	Tinco Paccha, Junín; Perú (**TP**)	B	16	5
	Tarma Tambo, Junín; Perú (**TT**)	B	8	8
	Sto. Domingo de Cachi Cachi, Junín; Perú (**CC**)	B	12	8
	San Pedro de Huacarpana, Ica; Perú (**HC**)	B	17	7
	Ayavi-Tambo-Huaytará, Huancavelica; Perú (**AY**)	B	7	8
	R. N. de Pampa Galeras, Ayacucho; Perú (**PG**)	B	21	15
	Cerro Azul, Cuzco; Perú (**CA**)	B	27	17
	S.A.I.S. Picotani, Puno; Perú (**PI**) (captivity)	B	10	7
	Ingenio Huacullani, Puno; Perú (**IG**)	B	16	16
	Parque Nacional Lauca; Chile (**LA**)	S, B	34	29
	Lagunillas; Chile (**LG**)	B	16	16
	Corral Ankara, Ankara; Chile (**AN**)	B	10	9
	Salar de Surire; Chile (**SS**)	B	17	17
Southern region		–	112	128
	Salar de Ascotán; Chile (**SA**)	F	17	28
	Paso Jama; Chile (**PJ**)	F	5	4
	P. N. Llullaillaco; Chile (**LL**)	B	9	10
	P. N. Nevado Tres Cruces; Chile (**TR**)	M, F, S	9	11
	Santa Catalina, Jujuy; Argentina (**SC**)	B	21	21
	INTA Abra Pampa, Jujuy; Argentina (**IN**) (captivity)	B	28	33
	Laguna Blanca, Catamarca; Argentina (**LB**)	B	20	18
	San Juan; Argentina (**SJ**)	B	3	3

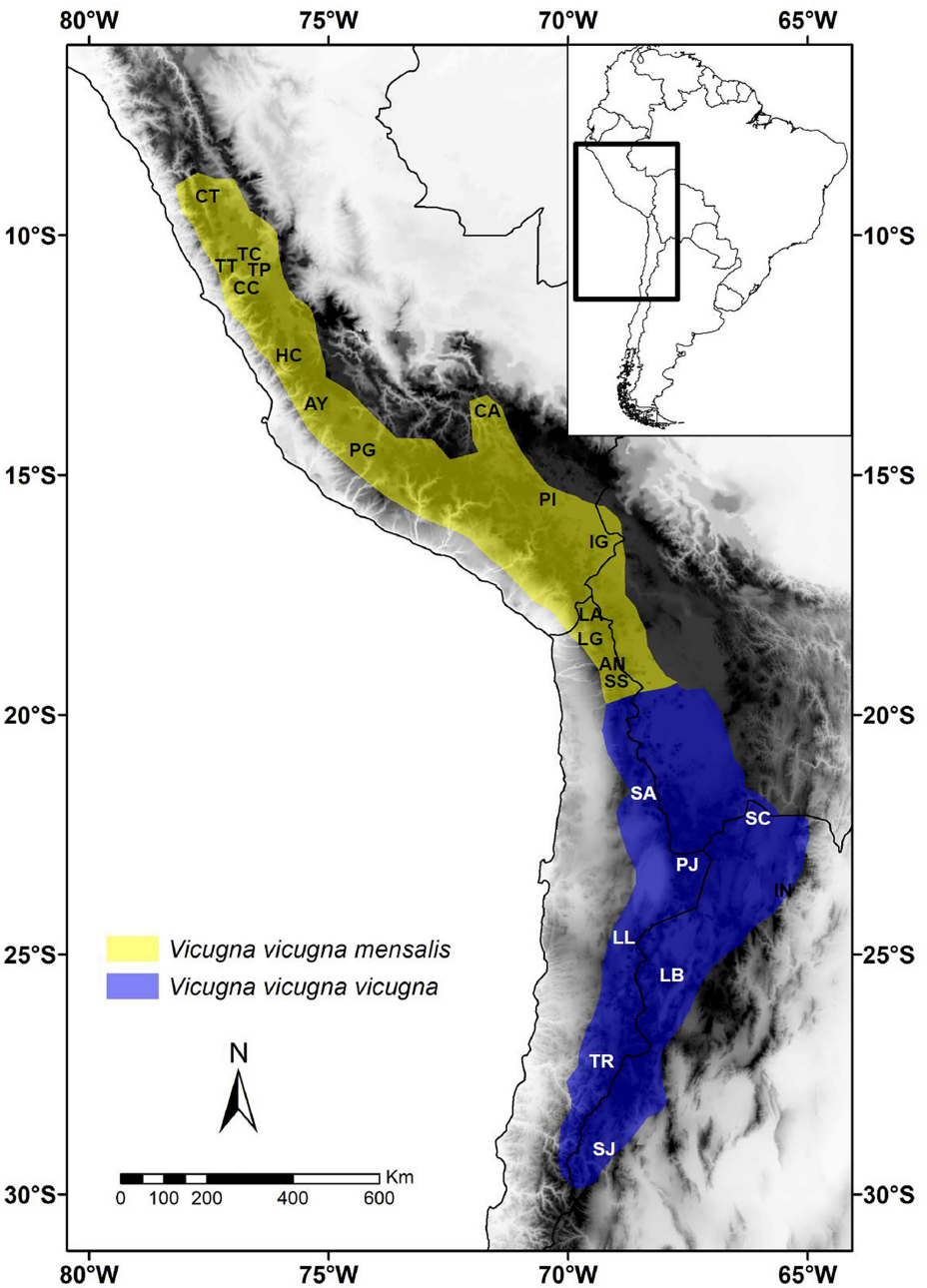


FIGURE 1 | Map of the geographic distribution of *Vicugna vicugna mensalis* (yellow) and *V. v. vicugna* (blue) and location of sampled populations in South America. Localities names are: Catac (CT), Tinco Cancha (TC), Tinco Paccha (TP), Tarma Tambo (TT), Sto. Domingo de Cachi Cachi (CC), San Pedro de Hacarpana (HC), Ayavi-Tambo-Huaytará (AY), R. N. Pampa Galeras (PG), Cerro Azul (CA), S.A.I.S. Picotani (PI), Ingenio Hacullani (IG), Parque Nacional Lauca (LA), Lagunillas (LG), Corral Ankara (AN), Salar de Surire (SS), Salar de Ascotán (SA), Paso Jama (PJ), P. N. Llullaillaco (LL), P. N. Nevado Tres Cruces (TR), Santa Catalina (SC), INTA Abra Pampa (IN), Laguna Blanca (LB), San Juan (SJ) (**Table 1**).

United States). DNA from skin and muscle samples was purified using proteinase K digestion and a standard phenol–chloroform protocol (Sambrook et al., 1989). DNA from feces was extracted using the QIAamp DNA Stool Mini Kit (QIAGEN, Valencia, CA, United States) in a separate non-genetic-oriented laboratory.

Mitochondrial DNA

Three hundred and eighty-five base pairs of the left domain of the mitochondrial control region (mtDNA-CR) was amplified using the camelid and vicuña-specific primers LthrArtio (5′-GGT CCT GTA AGC CGA AAA AGG A-3′), H15998V (5′-CCA GCT TCA ATT GAT TTG ACT GCG-3′), Loop007V

(5′-GTA CTA AAA GAG AAT TTT ATG TC-3′), H362 (5′-GGT TTC ACG CGG CAT GGT GAT T-3′) (Marín, 2004). Amplification was performed in 50 ml with ∼30 ng genomic DNA, 1x reaction buffer (8 mM Tris-HCl (pH 8.4), 20 mM KCl (Invitrogen, Gibco, Life Technologies, Invitrogen Ltd., Paisley, United Kingdom), 2 mM $MgCl_2$, 25 μM each of dNTP, 0.5 μM each primer and 0.1 U/μl Taq polymerase (Invitrogen, Gibco, Life Technologies). Thermocycling conditions were: 95°C for 10 min, followed by 30–35 cycles of 94°C for 45 s, 62°C for 45 s, 72°C for 45 s, then 72°C for 5 min. PCR products were purified using the GeneClean Turbo for PCR Kit (Bio101) following the manufacturer's instructions. Products were sequenced in forward and reverse directions using BigDye chemistry on an ABI Prism 3100 semiautomated DNA analyzer, and consensus sequences were generated and aligned using Geneious v.9.1.5 (Biomatters, Auckland, New Zealand). The final alignment was trimmed to 328 bp beginning at the 5′ left domain of the d-loop.

The number of segregating sites (*S*) and haplotypes (*nh*), haplotype diversity (*h*) (Nei, 1987), nucleotide diversity (π) and average number of nucleotide differences between pairs of sequences (*k*) were estimated using ARLEQUIN 3.5.1.2 (Excoffier and Lischer, 2010). A statistical parsimony network was constructed using TCS v1.21 (Clement et al., 2000) with default settings.

Microsatellite Markers

Fifteen autosomal dinucleotide microsatellite loci (YWLL08, YWLL29, YWLL36, YWLL38, YWLL40, YWLL43, YWLL46 – Lang et al., 1996, LCA5, LCA19, LCA22, LCA23 – Penedo et al., 1998, LCA65, LCA82 – Penedo et al., 1999, and LGU49, LGU68 – Sarno et al., 2000) were analyzed. Amplification was carried out in a 10 μL reaction volume, containing 50–100 ng of template DNA, 1.5–2.0 mM MgCl2, 0.325 μM of each primer, 0.2 mM dNTP, 1X polymerase chain reaction (PCR) buffer (QIAGEN) and 0.4 U Taq polymerase (QIAGEN). All PCR amplifications were performed in a PE9700 (Perkin Elmer Applied Biosystems) thermal cycler with cycling conditions of: initial denaturation at 95°C for 15 min, followed by 40 cycles of 95°C for 30 s, 52–57°C for 90 s and 72°C for 60 s, and a final extension of 72°C for 30 min. Amplification and genotyping of DNA from fecal samples was repeated two or three times. One primer of each pair was labeled with a fluorescent dye on the 5'-end, and fragments analyzed on an ABI-3100 sequencer (Perkin Elmer Applied Biosystems). Data collection, sizing of bands and analyses were carried out using Genemarker v. 1.70 (SoftGenetics).

We identified fecal samples that came from the same individual by searching for matching microsatellite genotypes using the Excel Microsatellite Toolkit (Park, 2001) and eliminated samples from the study if they showed more than 85% overlap. We also evaluated the existence of null alleles using the program Micro-Checker v. 2.2.3 (Van Oosterhout et al., 2004). ARLEQUIN 3.5.1.2 software (Excoffier and Lischer, 2010) was used to estimate allele frequency, observed heterozygosity (H_O), and expected heterozygosity (H_E). The inbreeding coefficient F_{IS} was estimated following Weir and Cockerham (1984) using FSTAT 2.9.4 (Goudet, 2005).

Genetic Structure and Gene Flow

We used the Bayesian clustering algorithm implemented in STRUCTURE v. 2.3.3 (Pritchard et al., 2000) to group the samples genotyped with microsatellites into *K* clusters. We tested values of *K* between 1 and 23, running STRUCTURE five times for each value of *K*, and using Evanno's Δ*K* method to determine the most suitable number of clusters (Evanno et al., 2005). STRUCTURE was run using the admixture model and correlated allele frequencies, as recommended for populations that are likely to be similar due to migration or shared ancestry (Falush et al., 2003; Pritchard et al., 2007). 500,000 iterations were used to estimate *K* after a burn-in period of 30,000 iterations.

Based on the STRUCTURE results we found $K = 2$ (**Figure 2**) to be the most suitable clustering solution (with each cluster corresponding to one subspecies; see results and

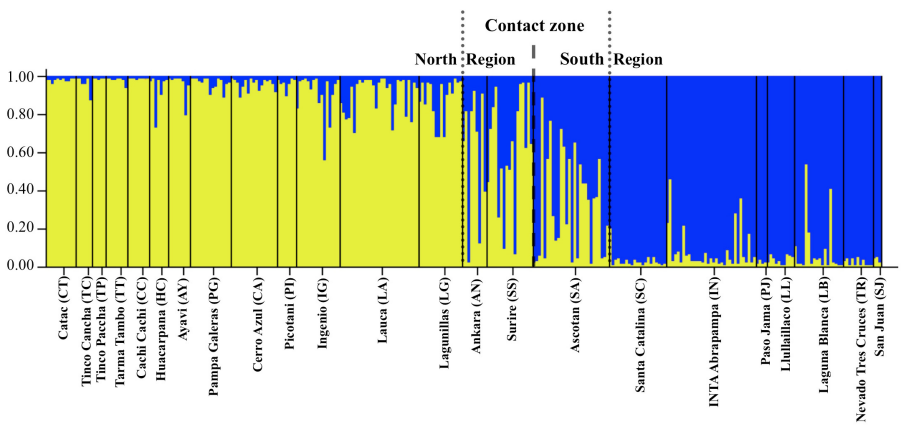


FIGURE 2 | Plot of posterior probability of assignment for 307 vicuñas (vertical lines) to two genetic clusters based on Bayesian analysis of variation at 15 microsatellite loci. Individuals are grouped by locality, and localities are indicated along the horizontal axis. Yellow, Genetic Cluster 1: North group; blue, Genetic Cluster 2: South group.

Figure 2). We used these results to selected the 70 least admixed individuals (35 from northern and 35 from southern localities) to simulate a hybrid population with HybridLab 1.0 (Nielsen et al., 2006). Using these three populations we assessed to which of them each sample in the dataset would be assigned. Furthermore, we also estimated the migration rate between *V. v. mensalis* and *V. v. vicugna* and the hybrid population using BayesAss 3.0 (Wilson and Rannala, 2003; Rannala, 2007). We assessed genetic differentiation between sampling localities using F_{ST} (Weir and Cockerham, 1984) estimated with the microsatellite data and mitochondrial DNA data in ARLEQUIN 3.5.1.2 (Excoffier and Lischer, 2010) with 10,000 permutations to assess significance. We also estimated pairwise population differentiation between sampling localities with Jost's D in GENODIVE v.2.0b22 (Meirmans and Van Tienderen, 2004), as this method is independent of the amount of within population diversity (Jost, 2008).

Spatial Autocorrelation Analyses

We tested for spatial autocorrelation in the data at various distance classes using Genalex 6.5 (Peakall and Smouse, 2012). We used Euclidean genetic and geographic distances between pairs of individuals for the species as a whole and at two additional hierarchical levels: (i) separated for males and females to test for sex-biased dispersal and (ii) separated by subspecies. Correlograms of *r*-values were estimated as a function of geographic distance using 200-km fixed distance class bins (shorter distances resulted in few observations per bin whereas longer distances compromised resolution of fine-scale genetic structure). We followed Banks and Peakall (2012) to test for correlogram significance and heterogeneity in allele autocorrelation between sex or subspecies groups using **Ω** and T2 statistics. One thousand bootstrap permutations were used to estimate the 95% confidence intervals around $r = 0$ (no correlation between genetic variation in individuals in the same bin) and to test if the observed and expected *r-values* were significantly different form each other. Lastly, we tested for the presence of isolation by distance (IBD) using Mantel tests between the matrix of geographic distances between localities and the matrix of Fst estimated for either the mtDNA or the microsatellite loci.

Migration–Drift Equilibrium

We tested the relative likelihood of a gene flow vs. a genetic drift model using the program 2MOD (Ciofi et al., 1999) (**Supplementary Figure 1**). Three different datasets were analyzed with 2MOD to test whether the inferred model was affected by the data used. The first dataset consisted of two populations, one made up of the individuals with the highest probability of belonging to *V. v. mensalis* and the other of the individuals with the highest probability of belonging to *V. v. vicugna* based on the STRUCTURE results for $K = 2$ and with a minimum Q threshold of 0.75 indicating the corresponding population ancestry. The second dataset consisted of one population composed of all individuals with a *Q*-value > 0.5 based on the STRUCTURE results for $K = 2$ and the other population made up of all the remaining individuals. The third dataset consisted of the data for each locality separately. Each analysis was run independently three times for 200,000 iterations of the Markov Chain Monte Carlo algorithm with the first 10% of the simulations discarded as burn-in.

Demographic History

The mtDNA of the northern and southern regions (one group per subspecies) were used to estimate Tajima's D and Fu's Fs statistics (Tajima, 1989; Fu, 1997) with 10,000 simulations to assess significance in ARLEQUIN (Schneider and Excoffier, 1999). These analyses were complemented with Bayesian Skyline Plot (BSP) analyses in BEAST run separately for the individuals of each subspecies (one group per subspecies; Bouckaert et al., 2014). BEAST's Markov Chain Monte Carlo algorithm was run independently three times for 50 million steps discarding the first 10% as burn-in and until ESS values above 200 would be obtained. A substitution rate of 1.2%/million years was used to scale the BSP in years (as in Almathen et al., 2016).

To identify demographic scenarios that explain current diversity patterns in the northern and southern regions, we used the coalescent-based framework implemented in MSVAR v1.3 (Storz and Beaumont, 2002) using microsatellite loci of each sampling locality (see section "Results"). MSVAR estimates the recent effective population size (N0), the ancestral effective population size (Nt), and the time (t) at which the effective population size changed from Nt to N0. Three independent runs of MSVAR were carried out including wide prior distributions of the model parameters and accounting for the possibility that the populations remained stable over time ($Nt \sim N0$), that there was a bottleneck ($Nt > N0$), or a population expansion ($Nt < N0$). Prior distributions are log-normal distribution parameterized with the mean and standard deviation for each parameter and truncated at zero following Storz and Beaumont (2002) (**Supplementary Table 3**). Because there is no vicuña specific microsatellite mutation rate available, we used a range of typical vertebrate microsatellite mutation rates varying between $10^{-2.5}$ and $10^{-4.5}$ (Weber and Wong, 1993; Di Rienzo et al., 1994; Brohede et al., 2002; Bulut et al., 2009) (**Supplementary Table 3**). MSVAR was run for a total of 400 million iterations under each demographic model discarding the initial 20% of the MCMC steps as burn-in. The independent runs were used to estimate the mode of the posterior distributions of each parameter (N0, Nt, and t) and their corresponding 90% highest posterior density interval. A generation length of 3 years (Franklin, 1983; González et al., 2006) was used to rescale the t parameter in years. Convergence of the runs was estimated with the Gelman and Rubin's diagnostic using the CODA library (Plummer et al., 2006) in R (R Development Core Team, 2009).

RESULTS

Genetic Diversity

Among the 353 samples genotyped with microsatellites we detected on average 13 alleles/microsatellite. The number of alleles per locus ranged from 5 to 23, and 38 alleles were unique to a single locality. No deviation from H-W equilibrium was found

due to an excess of homozygotes, however, a significant excess of heterozygotes ($P < 0.0151$, FDR adjusted critical value) was for various markers in different populations, with ten populations showing no deviation from HWE for any locus. The remaining population showed between one and a maximum of eight locus not in HWE, with an overall mean of 2 loci per locality not in HWE. Overall no linkage disequilibrium was found in the sampling localities, with the exception of 12% o the microsatellite pairwise comparisons in Lauca, and one microsatellite pairwise comparison in Lagunillas, another one in Ascotán, and two in Santa Catalina (however, the loci in these comparisons in the last three populations were not the same; FDR adjusted $P < 0.00955$). Estimates of genetic diversity excluding these loci were not significantly different from those estimated with all loci, thus, we kept these markers for further analyses (Welch corrected *t*-test - *p*-value > 0.05). We found consistently moderate to high levels of genetic diversity (mean expected heterozygosity ranged from 0.45 to 0.78) and high values for allelic richness (mean RA ranged from 2.67 to 7.53) (**Table 2**) relative to other South American mammals, e.g., andean bear (Ruiz-Garcia et al., 2005), guanacos (Marín et al., 2013a), guigna (Napolitano et al., 2015). In the mtDNA we found 52 variable positions (17.33%), 34 transitions, 7 transversions and one insertion among 385 nucleotides, resulting in 57 haplotypes ($h = 0.794$) among 376 partial sequences of the 5′ end of the control region. Among variable sites, 37 (71.15%) were parsimonious informative. Haplotype (h) and nucleotide diversity (π) ranged between 0–0.92 and 0–0.35 (**Table 2**). GenBank accession for the publicly available data are AY535173–AY535284 and KY420493–KY420569 for the newly generated sequences.

Genetic Structure and Gene Flow

Results of the STRUCTURE analysis indicated that the best clustering solution was $K = 2$ based on the ΔK method (**Supplementary Figures 2A,B**) with most of the samples from the Northern region being assigned to one cluster and most of the samples from the Southern region being assigned to the other cluster. Of the 353 individuals, 269 had a clear predominant heritage with Q-values > 0.75. Of the 173 individuals sampled in the Northern region, 155 presented a predominantly northern genetic background, while 20 presented a mixed South – North origin and 4 had clear South genetic heritage with Q-values indicating southern ancestry larger than 0.75 (**Table 3**). Of the 108 individuals sampled in the Southern Region, 95 presented a predominantly southern genetic background ($Q > 0.75$), while 13 were of mixed origin, and 2 were assigned to the Northern cluster (i.e., their Q-values indicating a northern ancestry were > 0.75). Overall, most

TABLE 2 | Genetic diversity indices from 15 microsatellite loci and mtDNA Control Region sequences by localities (defined in **Table 1**).

	Microsatellites			mtDNA			
Region Localities	***A* ± SD**	***Ho* ± SD**	***He* ± SD**	***n***	***np***	***h* ± SD**	**π ± SD**
North	10.73 ± 4.96	0.40 ± 0.16	0.70 ± 0.18	30	26	0.71 ± 0.03	0.010 ± 0.001
CT	2.67 ± 1.03	0.39 ± 0.25	0.45 ± 0.12	3	1	0.60 ± 0.08	0.008 ± 0.001
TC	3.54 ± 1.20	0.39 ± 0.26	0.60 ± 0.26	2	0	0.40 ± 0.11	0.005 ± 0.002
TP	3.33 ± 1.18	0.62 ± 0.34	0.65 ± 0.20	2	1	0.23 ± 0.13	0.001 ± 0.000
TT	3.42 ± 1.56	0.36 ± 0.22	0.56 ± 0.21	3	0	0.71 ± 0.12	0.003 ± 0.001
CC	3.08 ± 1.17	0.44 ± 0.25	0.57 ± 0.14	2	0	0.17 ± 0.13	0.001 ± 0.001
HC	4.23 ± 2.39	0.46 ± 0.28	0.66 ± 0.18	4	1	0.63 ± 0.08	0.009 ± 0.001
AY	3.92 ± 1.32	0.42 ± 0.25	0.63 ± 0.16	2	0	0.57 ± 0.12	0.002 ± 0.000
PG	5.07 ± 2.25	0.50 ± 0.22	0.64 ± 0.21	7	3	0.82 ± 0.05	0.010 ± 0.001
CA	5.21 ± 2.78	0.42 ± 0.27	0.57 ± 0.27	4	1	0.33 ± 0.11	0.001 ± 0.001
PI (captivity)	3.39 ± 1.33	0.48 ± 0.25	0.60 ± 0.18	1	0	0.00 ± 0.00	0.000 ± 0.000
IG	5.73 ± 2.58	0.46 ± 0.23	0.64 ± 0.24	3	0	0.71 ± 0.05	0.003 ± 0.000
LA	7.53 ± 3.58	0.38 ± 0.23	0.68 ± 0.19	10	5	0.73 ± 0.07	0.013 ± 0.004
LG	5.20 ± 2.37	0.42 ± 0.15	0.64 ± 0.19	6	3	0.62 ± 0.14	0.006 ± 0.003
AN	4.53 ± 1.77	0.46 ± 0.24	0.68 ± 0.19	4	1	0.64 ± 0.15	0.004 ± 0.001
SS	6.33 ± 2.52	0.40 ± 0.27	0.71 ± 0.18	8	4	0.89 ± 0.05	0.027 ± 0.002
South	11.33 ± 4.50	0.37 ± 0.16	0.74 ± 0.18	31	27	0.84 ± 0.02	0.028 ± 0.001
SA	7.00 ± 3.40	0.35 ± 0.22	0.66 ± 0.23	7	2	0.79 ± 0.08	0.019 ± 0.003
SC	6.40 ± 2.06	0.36 ± 0.19	0.70 ± 0.14	7	3	0.76 ± 0.07	0.026 ± 0.002
IN (captivity)	6.33 ± 2.19	0.34 ± 0.21	0.66 ± 0.19	4	1	0.66 ± 0.05	0.025 ± 0.002
PJ	3.56 ± 1.24	0.44 ± 0.41	0.78 ± 0.16	3	0	0.70 ± 0.22	0.029 ± 0.010
LL	4.36 ± 1.45	0.38 ± 0.23	0.65 ± 0.14	6	1	0.92 ± 0.07	0.031 ± 0.004
LB	5.13 ± 2.23	0.33 ± 0.23	0.66 ± 0.15	9	3	0.83 ± 0.07	0.030 ± 0.002
TR	5.27 ± 1.83	0.49 ± 0.27	0.70 ± 0.19	5	2	0.81 ± 0.12	0.035 ± 0.005
SJ	2.69 ± 0.63	0.49 ± 0.32	0.66 ± 0.14	2	1	0.67 ± 0.31	0.018 ± 0.008

A, number of alleles per locus; He, expected heterozygosity; Ho, observed heterozygosity; n, number of haplotypes; np, number of private haplotypes; h, haplotype diversity; π, nucleotide diversity; SD, standard deviation.

TABLE 3 | Percentage of individuals assigned to the North or South genetic clusters with nuclear DNA ($Q > 75\%$, STRUCTURE analysis).

Localities	Microsatellites		
	North	Hybrids	South
CT	100.0 (11)	0.0 (0)	0.0 (0)
TC	100.0 (6)	0.0 (0)	0.0 (0)
TP	100.0 (5)	0.0 (0)	0.0 (0)
TT	100.0 (8)	0.0 (0)	0.0 (0)
CC	100.0 (8)	0.0 (0)	0.0 (0)
HC	85.7 (6)	14.3 (1)	0.0 (0)
AY	100.0 (8)	0.0 (0)	0.0 (0)
PG	100.0 (15)	0.0 (0)	0.0 (0)
CA	100.0 (17)	0.0 (0)	0.0 (0)
PI	100.0 (7)	0.0 (0)	0.0 (0)
IG	87.5 (14)	12.5 (2)	0.0 (0)
LA	93.1 (27)	6.9 (2)	0.0 (0)
LG	81.3 (13)	18.7 (3)	0.0 (0)
AN	**44.4 (4)**	**33.3 (3)**	**22.2 (2)**
SS	**35.3 (6)**	**52.9 (9)**	**11.8 (2)**
SA	**7.1 (2)**	**46.4 (13)**	**46.4 (13)**
SC	0.0 (0)	0.0 (0)	100.0 (21)
IN	0.0 (0)	9.1 (3)	90.9 (30)
PJ	0.0 (0)	0.0 (0)	100.0 (4)
LL	0.0 (0)	0.0 (0)	100.0 (10)
LB	0.0 (0)	11.1 (2)	88.9 (16)
TR	0.0 (0)	0.0 (0)	100.0 (11)
SJ	0.0 (0)	0.0 (0)	100.0 (3)

The number of individuals from each locality is in parentheses. Localities containing animals from of both clusters, considered to be hybrids ($Q > 25\%$), are shaded and in bold.

of the individuals of mixed or incongruent heritage in both clusters were sampled in the localities of AN, SS and SA, which correspond to the contact zone between the North and South regions (**Figure 2** and **Table 3**).

The pairwise analysis of divergence between sampling localities showed significant genetic differentiation between ∼58% of the localities pairwise comparisons measured with the F_{ST} and *(D)phi-st* (**Supplementary Table 1**). Significant pairwise differentiation estimated with F_{ST} ranged from low (0.037) to high values (0.612), with the greatest divergence observed between the CC and PI localities. *D* pairwise divergence values were larger than the F_{ST} values with significant values ranging between 0.19 and 0.95 and the highest divergence observed between the AY and PI localities. Furthermore, we also found a significant negative correlation (−0.56; *P ∼2.2e-16*) in average heterozygosity and pairwise F_{ST} between sampling localities suggesting that the observed divergence between localities may be driven by genetic drift rather than the build up of unique mutations present at the different localities (Weeks et al., 2016).

When dividing the samples into two groups representing the two clusters identified by STRUCTURE reflecting the two subspecies, the F_{ST} including and excluding hybrids were 0.0779 and 0.0998, respectively, and both were significant ($p < 0.005$). A division between the North and South regions was also observed with the haplotype network calculated with the 57 mtDNA Control Region haplotypes (**Figure 3**). Among these haplotypes 28 were found in *V. v. mensalis* (North) and 25 were found only in *V. v. vicugna* (South), while four where shared between the two subspecies and thus found in both the North and South Regions (haplotype 2, 6, 17, and 21). Additionally, haplotypes 27, 29, and 30 found only on *V. v. mensalis* grouped together with those of *V. v. vicugna*, while haplotypes 33 and 34 found only on *V. v. vicugna* clustered with to the *V. v. mensalis* haplotypes. The 57 haplotypes are connected with a maximum of 48 mutational steps, with most genetic variation localized regionally and the main link between the two geographic regions of the network diverging by five mutations. Consistent with the STRUCTURE results, the haplotypes shared between the two regions occurred in the sampling localities in the contact zone (AN, SS, and SA).

Spatial Autocorrelation Analysis

We found significant spatial autocorrelation among individuals for the species as whole ($\Omega = 96.9 = 7$, $p < 0.001$) and at both hierarchical levels, implying dispersal is limited at the spatial scale, with greater resemblance between individuals at shorter distances and decreasing resemblance as distance increases. Although both males and females presented a spatial autocorrelation pattern indicative of isolation by distance, these were significantly different from each other ($\Omega = 62.3$, $p < 0.001$) with females presenting almost twice as much similarity than males at the smaller distance class (T2 = 63.5, $p < 0.001$ – 200 km distance class). None the less, as geographical distance increases the differences between *r*-values in both sexes almost disappears (**Figures 4A,B**). We also found significant differences in autocorrelation between subspecies ($\Omega = 36.2$, $p < 0.001$; **Figures 4C,D**) with the northern vicuña showing a stronger effect of geographic distance on their similarity than in the southern vicuña which shows approximately the same similarity across all distance classes. The analysis of IBD using Mantel tests with the mtDNA data resulted in a significant correlation between the matrix of geographic distances between localities and the Fst between localities ($r \sim 0.36$, *p*-value = 0.00016). Equivalent results were obtained for this analysis using the microsatellite data ($r \sim 0.38$, *p*-value = 0.00001).

Migration–Drift Equilibrium

Results of the coalescent analyses to tests for gene flow + genetic drift vs. a genetic drift only model showed that all simulations support a genetic drift + gene flow model as an explanation of the observed genetic variability (the posterior probability of the model of genetic drift + gene flow was higher than 0.98 for all tests; **Supplementary Table 5**). This result is the same for all three alternative dataset configurations tested and indicates that vicuña localities have historically not been isolated from each other but, rather, inter connected.

Demographic History

Based on the mtDNA data, the North Region showed negative Fu's *F*-values ($F_s = -2.3279$, $P < 0.05$) and Tajima's *D* ($D = -1.7341$, $P > 0.05$) consistent with a pattern of population

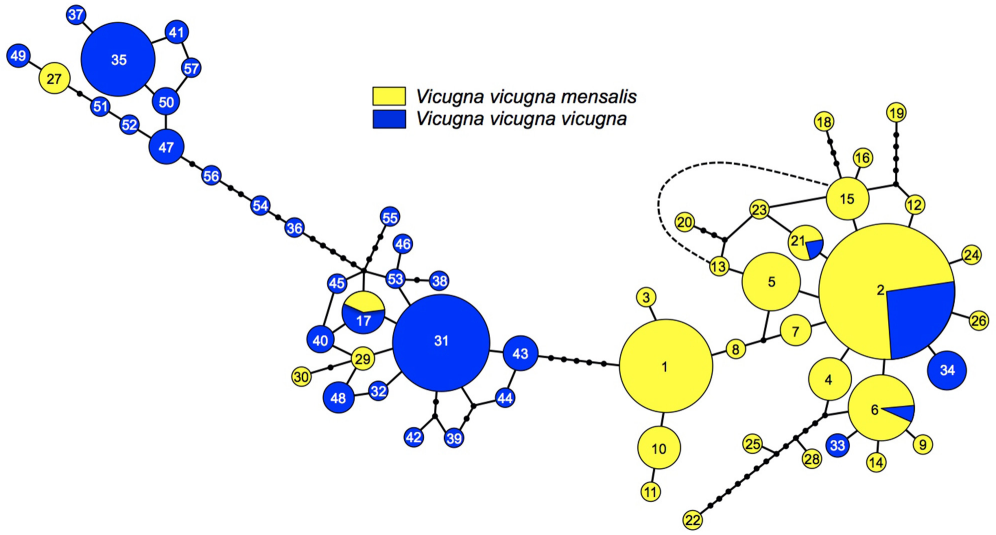


FIGURE 3 | Minimum spanning network representing the relationships among 57 control region haplotypes, numbers represent each haplotype (see **Supplementary Table 2**). Circle sizes correspond to haplotype frequencies. Branches without circles correspond to one difference between haplotypes, and each small black circle corresponds to one additional mutation. Dashed line represents one mutational step between haplotypes 13 and 16.

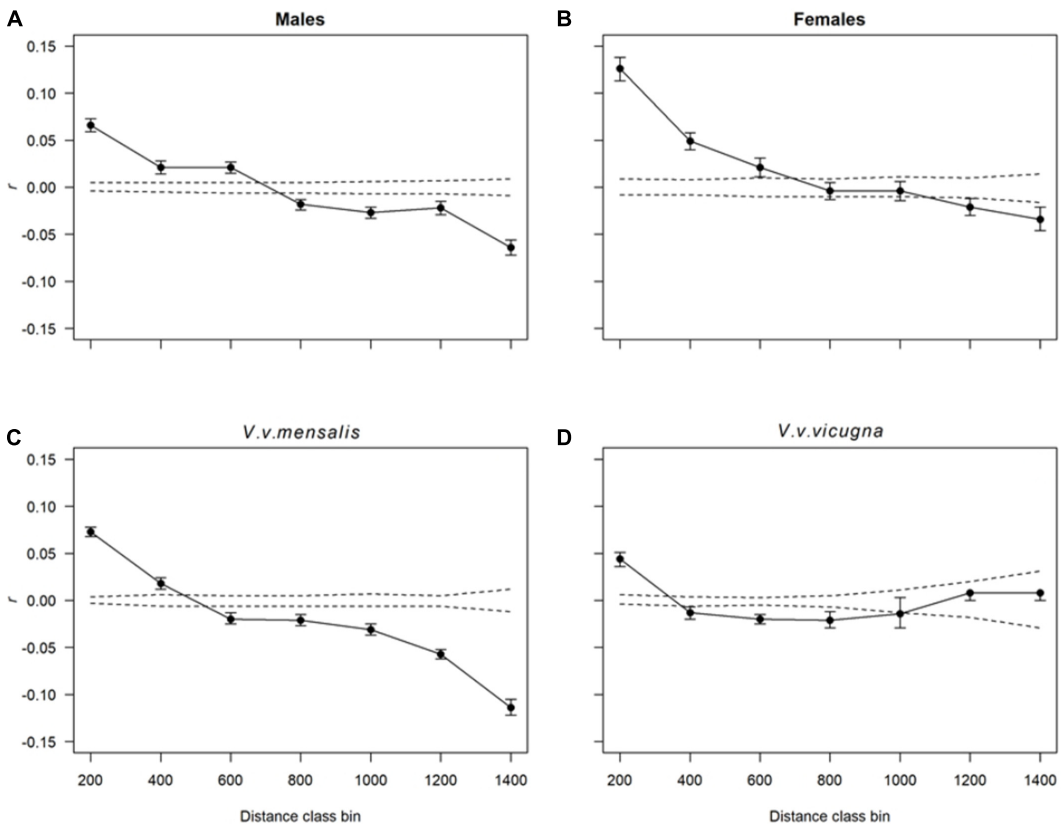


FIGURE 4 | Correlograms showing the combined spatial correlation *r* across transects as a function of distance. Dotted lines correspond to the 95% CI about the null hypothesis of a random distribution of genetic variation over space (i.e., no effect of geographic distance on *r*). Each *r* value has 95% confidence error bars shown. Distance class bins are shown in kilometers. Each plot shows the autocorrelation for distance class sizes of 200 km estimated for **(A)** males, **(B)** females, **(C)** all northern vicuñas (*V. v. mensalis*) and **(D)** southern vicuñas (*V. v. vicugna*).

expansion, however, Tajima's D was not significant. For the South region both tests were not significant ($F_s = -0.0863$, $P > 0.10$; $D = 1.2433$, $P > 0.05$) suggesting a stable population history. These results are consistent with the presence of a major haplotype in the northern region (Hap2), while in the southern region two distantly related haplotypes occur at moderate to high frequency (Hap31 and 35) with many low frequency haplotypes in between them. The BSP analysis for each subspecies clearly shows a higher effective population size for *V. v. mensalis* (**Figures 5A,B**). This same analysis also supports a recent population expansion for both *V. v. mensalis* and *V. v. vicugna* starting approximately 3,000 years ago (**Figures 5A,B**). However, *V. v. mensalis* presents a population decrease starting approximately 800 years ago (**Figure 5A**).

The demographic analysis using the microsatellite data and MsVar showed consistent results under all three demographic

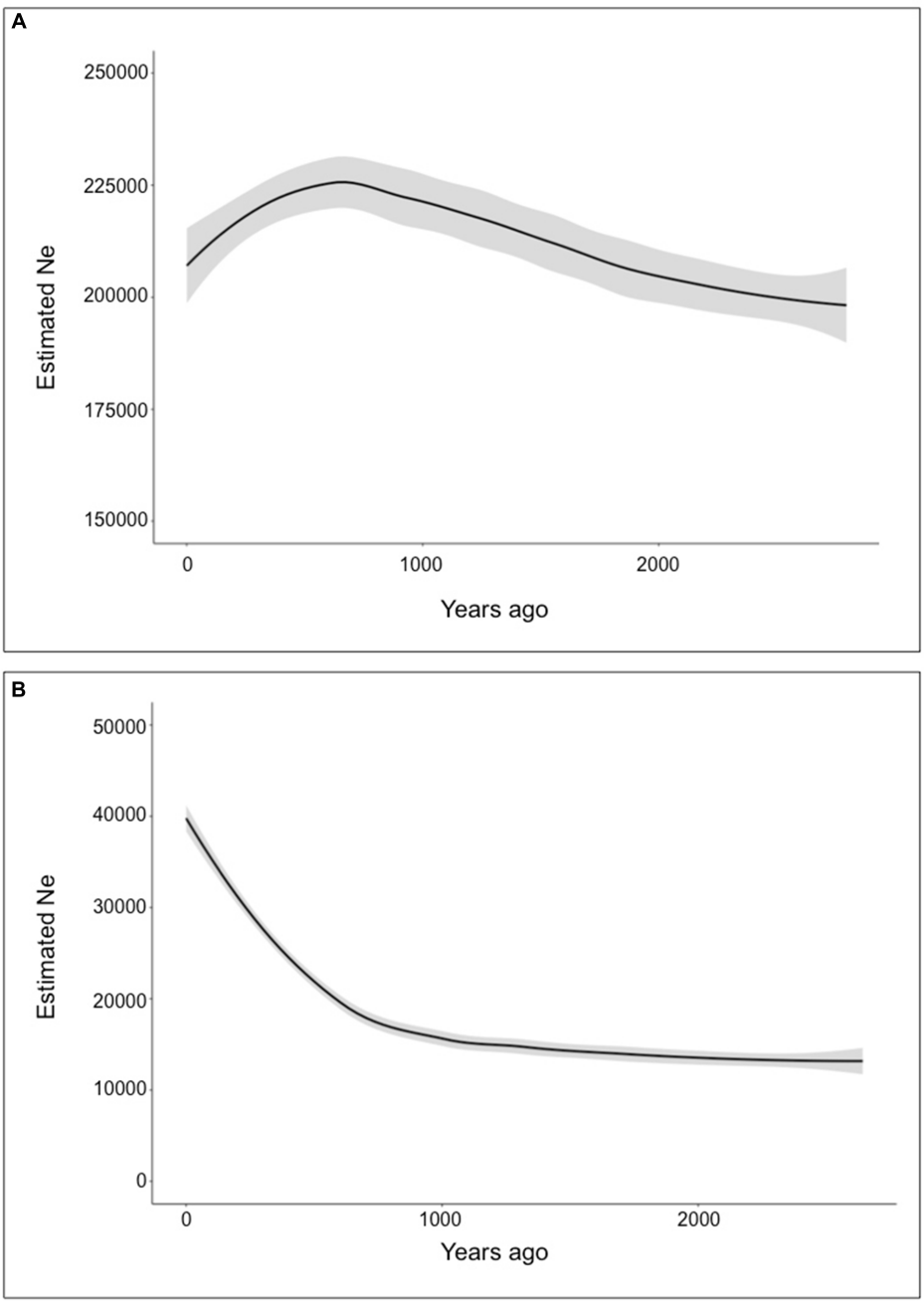


FIGURE 5 | Demographic analysis of *Vicugna vicugna* using the Bayesian Skyline Plot Method. Gray background represents error margins. **(A)** Bayesian Skyline Plot of *Vicugna vicugna mensalis*. **(B)** Bayesian Skyline Plot of *Vicugna vicugna vicugna*.

models tested, i.e., (1) no demographic change, (2) bottleneck and (3) population expansion. The three independent runs for each sampling locality presented Gelman and Rubin's statistics lower than 1.2. In all cases MSVAR detected evidence for major effective population size decline at all localities, consistent with current or recent small census sizes (**Figure 6** and **Supplementary Table 3**). All localities analyzed presented large ancestral effective population sizes on the order of ∼22,000 individuals [with 95% highest posterior density intervals (HPD) between ∼5,000 and ∼100,000; **Supplementary Table 3**]. Across localities the time of the bottleneck was on average ∼7,600 years before the present (YBP; HPD ∼760–125,000 YBP). Following this event, the effective population size in vicuña reached on average less than 1,000 (HPD ∼200–∼22,000; **Supplementary Table 4**). Similar ancestral effective population size and bottleneck dates, combined with the results of migration–drift and isolation by distance analyses, suggests that current vicuñas descend from a single large ancestral population that only recently started diverging probably through genetic drift.

DISCUSSION

Here, we present the most comprehensive analysis to date of genetic variation in vicuña across the species range. The molecular markers used present contrasting patterns regarding vicuña evolutionary history with conflicting evidence regarding the split of the species into two different taxonomic units. For instance, analysis of the mtDNA haplotypes largely supports a split into two vicuña groups, with each group dominated by haplotypes specific to the northern (i.e., *V. v. mensalis*) or the southern (i.e., *V. v. vicugna*) groups, respectively. This apparent differentiation is supported by the divergence analysis that results in a large and significant Φ_{CT} of 0.4. Nevertheless, some haplotypes are shared between the two groups of vicuñas, particularly at localities in the contact zone between 18°S (Ankara, AN) and 21°S (Salar de Ascotan, SA). The data presented here evidences gene flow between the two vicuña types as reflected by haplotype sharing between animals of each type (e.g., haplotypes 2, 6, 17, and 21) deriving from localities at or near the contact zone between the two types (Salar de Ascotan, Santa Catalina, Lauca, Salar de Surire). This is further supported the observation of phenotypically *mensalis* individuals that carry *vicugna* haplotypes (i.e., seven individuals from Salar de Surire carrying haplotypes 27 (four animals), 29 (2 animals), and 30 (1 animal) and conversely, phenotypically *vicugna* individuals that carry *mensalis* haplotypes (i.e., two individuals from Salar de Ascotan carrying haplotype 33 and one carrying haplotype 34, and four individuals from Santa Catalina carrying haplotype 33).

STRUCTURE analysis of the microsatellite data also identifies the presence of two clusters in the dataset (K = 2), and supports a contact zone between the two clusters where animals in that region present composite genotypes at intermediate allele frequencies to those observed in more northern or southern vicuñas (**Figure 2**). Our dataset did not include animals from Bolivia, a region in the contact zone between the two clusters identified here. However, we expect that considering Bolivia's location, vicuña samples from that part of the range might belong to the hybrid set of genotypes detected here for the contact zone. Such a pattern would explain the lack of taxonomic differentiation previously observed in Bolivia (Sarno et al., 2004). Interestingly, the contact zone between the two vicuña distribution ranges broadly coincides with the area occupied by the Tauka palaeolake which formed after the last glacial maximum and disappeared around 8,500 years ago (Blard et al., 2011). At its maximum, 16,000–14,500 BP, Tauka palaeolake is thought to have covered more than 52,000 km^2 (i.e., ∼75% the size of Lake Victoria in Africa) including the extant Lake Poopó and the Coipasa and Uyuni saltpans. However, following its disappearance, a narrow area between the mountains to the east of the

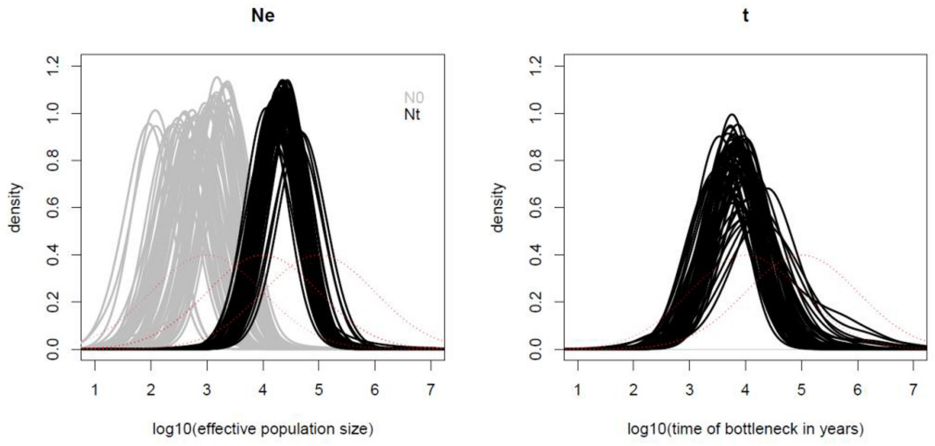


FIGURE 6 | Demographic analysis of *Vicugna vicugna* with MsVar. Panel **(A)** shows the posterior distributions of the current effective population size (gray lines), and ancestral effective population size (black lines). The time of the bottleneck (black lines) is shown in panel **(B)** for each of the three MSVAR replicates for each locality analyzed. The *x*-axis values are in log scale (e.g., 2 means 10^2). The red lines show the prior distributions.

Atacama Desert and the Uyuni saltpan opened enabling contact between the two vicuña groups, which otherwise would have been separated by the presence of this large lake in the Andean altiplano.

The best clustering solution identified by STRUCTURE with the microsatellite data, while reminiscent of the mtDNA split between the two vicuña subspecies, is more likely the outcome of isolation by distance (see Mantel test results and spatial autocorrelation) where individuals from localities at one end of the range distribution are more likely to resemble each other than individuals on the opposite end of the range. In such a scenario, instead of two populations being present, the data represents a single population with a gradient of intermediate genotypes between two contrasting extremes, as observed in **Figure 2**. Consistent with such pattern, individuals from the localities in the contact zone between the North and South ranges (i.e., Ankara, Surire and Ascotan), present a variety of genotypes ranging from mostly belonging to one of the two groups to almost presenting 50% ancestry from each group, while animals beyond this region are mostly of one genetic background. This is further supported by the phylogeographic analysis of the DBY gene of the Y chromosome which found no evidence for differences between *V. v. mensalis* y *V. v. vicugna* (Marín et al., 2017).

The change in genetic similarity between animals at increased distance was also observed with the spatial autocorrelation analyses. While splitting the dataset by sex resulted in both groups showing approximately the same isolation by distance pattern, females were more similar to each other at smaller distances (e.g., 200–400 km) than males. This pattern is consistent with females behaving more philopatric than males, with the later leaving their family groups upon becoming yearlings and form non-territorial bachelor groups which frequently have to move because of conflicts with local males with established territories (Koford, 1957; Franklin, 1983; Arzamendia et al., 2018). Yet, both females and males contributed similarly to gene flow at distance classes from 600 to 800 km, suggesting that beyond 800 km the effect of gene flow is limited, as few animals (if any) move that far.

A different pattern of spatial autocorrelation was observed when separately analyzing the northern and southern vicuñas. While northern vicuñas show the same isolation by distance pattern discussed above (**Figure 4**), the autocorrelogram for the southern vicuña drops quickly between 200 and 400 km, then seems to level off with similar *r*-values across further distances. Northern vicuñas inhabit a geographic range with higher habitat productivity and wider dietary resource availability than southern vicuñas. Thus, while northern vicuñas can find food resources in relative proximity (Franklin, 1982), southern vicuñas need to move over longer distances to find them, thereby increasing the probability of reproduction with animals that otherwise would be too far away (Arzamendia et al., 2018). However, such difference can also be achieved by populations with different levels of genetic variation, where populations with lower genetic diversity will experience a stronger effect of geographic distance if gene flow is low (as in northern vicuñas), while populations with a higher genetic diversity (as in southern vicuñas) need of a stronger reduction in gene flow to result in the same spatial pattern.

Extant vicuña populations are assigned to one of the two recognized vicuña subspecies; however, while these may present some morphological differences possibly reflecting local adaptation, their genetic variation suggests they form two extremes of a genetic continuum. Further evidence about the joint evolution of the two vicuña groups is provided by demographic modeling of the history of the various localities analyzed here, and assessment of whether these localities evolved independently from each other or connected via gene flow. The demographic analysis with MSVAR found that the extant vicuña genetic variation is the outcome of strong bottleneck that occurred ∼7,600 YBP (HPD ∼760–125,000 YBP). However, what is remarkable, is not only that all extant populations seem to have passed through this bottleneck at approximately same time, they all had a very similar ancestral effective population size (i.e., ∼25,000–HPD ∼5,000–100,000). It is likely that a single large vicuña population occupied a wide range across the Andes prior to a relatively recent bottleneck that dramatically reduced the effective population size to less than 1,000 (HPD ∼50–10,000). The main consequence of this event was fragmentation and isolation of previously well connected vicuña populations within their present distribution area (9°S to 29° S) and the small effective population size of pocket populations that survived. This hypothesis is supported by comparison of the model of evolution under genetic drift against a model that also included gene flow and which unambiguously showed that the latter better explains the extant genetic variation. While this analysis does not indicate modern connectivity between these localities, as has been previously suggested for vicuña (Casey et al., 2018), it supports that connection between them has been a major factor in the recent evolutionary history of *V. vicugna*.

The average bottleneck estimate across the vicuña populations is ∼7,600 YBP, but the range of variation from ∼700–125,000 YBP (**Supplementary Table 4**) reflect the uncertainty associated with estimates like generation length and mutation rate. Thus, while it is tempting to try to associate the inferred bottleneck with a particular event during the South American Holocene, it is safer to assume that it occurred sometime over the last 12,000 years. This period has been marked by dramatic changes across South America including the establishment of human hunters in the Peruvian high Andes ∼9,000 YBP (Aldenderfer, 1999) who, by ∼6,000 YBP, specialized on vicuña and guanaco (Wheeler et al., 1976). Additionally, this period of time included the transition from hunting to herding, with domestication of vicuña ∼6,000–5,500 YBP (Wheeler, 1995). It is possible that any or all of these events contributed to the demographic signal observed here, however, we are not able to pinpoint a single event. While the consistent results obtained across vicuña populations are indicative of the robustness of the genetic signature of the demographic change (Chikhi et al., 2010; Peter et al., 2010), future studies should be carried out using larger datasets (e.g., genome-wide polymorphisms) and other methodologies that are likely to

result in narrower confidence intervals of parameters of interest (e.g., approximate Bayesian computation).

Conservation Implications

Management Units

The evolutionary history of extant vicuñas is not at odds with the observation of morphological differences between animals across its range. In fact, it indicates that despite environmental changes during the Late Pleistocene and Holocene, *V. vicugna* maintained its genetic and taxonomic identity through time. Moreover, this identity remained despite the human population expansion in South America (<12,000 YBP) and their specialization in hunting vicuña (and guanaco) (Wheeler et al., 1976), as well as domestication of the vicuña at 6,000–5,500 YBP (Baid and Wheeler, 1993; Wheeler, 1995, 2012). Morphological variation across vicuña is likely to reflect the extensive territory they occupy and the different ecologies they are exposed to. Hence, morphological differences between the northern and southern groups would conform to ecotypes, as in other species, even if those where differences are not obviously reflected by the molecular markers used here (Courtois et al., 2003).

Establishing MUs is difficult because of differentiation by distance and the influence of genetics pools at sampled localities. MU determination depends on geographic areas with independent demographic dynamics between populations, whose individuals present a well-defined genetic structure and low migration rates (Marín et al., 2013a; Sveegaard et al., 2015; Yannic et al., 2016). At the large continental scale, it was not possible for us to identify discrete genetic clusters differentiating localities along the vicuña distribution range in this study, as it was in the study of Peruvian vicuña populations (Wheeler et al., 2001); therefore we propose the use of "continuous" MUs for the species. The main reason for this is that vicuña populations are defined by distance instead of by population discrete fragmentation or geographical barriers. By implementing this approach it is possible to include the spatial correlation information in defining the management area dimensions (e.g., 0–200 km) and protective actions for each locality (e.g., Caughley, 1994), which would enable extending the proposed MUs of Wheeler beyond the only four groups identified in their work (Wheeler et al., 2001).

Captive Populations

Two captive populations were included in this study Abra Pampa (Argentina) and Picotani (Peru). These populations present lower genetic variation as measured with both types of molecular markers than their immediate wild neighboring populations (i.e., Santa Catalina and Ingenio, respectively). The Abra Pampa Experimental Station has had captive vicuñas since 1933 (see Mosa, 1987). Although today's captive population is ∼1,200 individuals (Boswall, 1972; Mosa, 1987; Canedi and Pasini, 1996; Vicuña Convention, 2017) it has been reported that the founding population may have been as few as 22 animals (Canedi and Pasini, 1996). Our results suggest that this population has lower allelic richness and haplotypic diversity indexes than the other localities, probably as a consequence of a founder effect. None the less, the Abra Pampa confinement system has not resulted in a substantial loss of heterozygosity, supporting the hypothesis that a constant but low flow of wild vicuña into the captive herd has taken place (Anello et al., 2016). On the other hand, the Picotani animals have been in captivity since 1997 and they are completely enclosed and there is no breeding with wild vicuña. Our results indicate that the majority of genetic diversity estimators show a reduction in genetic variation in this captive population (e.g., only one mtDNA haplotype was found in comparison to three in neighboring Ingenio) probably due to the founder effect and despite of the larger starting population relative to Abra Pampa. The poor mtDNA genetic variation in captive animals from Picotani is worrying if some management decisions are taken at short-term such as releasing into the wild or translocating for repopulation or productive purposes, therefore a genetic impact assessment is urgently needed for decision support. These results help both setting a basal line for monitoring genetic diversity in these captive populations, but also provide information relevant for the development of an improved long-term captive management strategy at both locations to mitigate the observed loss of genetic variation.

CONCLUSION

Here, we present the most extensive genetic analysis of *Vicugna vicugna* to date. These results suggest that the two morphological variants of vicuña, i.e., the northern *V. v. mensalis* and the southern *V. v. vicugna*, were until recently closely interconnected with each other, or probably part of a single large population that passed through a strong bottleneck that left small isolated populations across a vast geographic range. Furthermore, extant vicuña genetic variation is better explained by a model of isolation by distance rather than by two discrete populations. However, given the extent of the vicuña geographic range and variation in the environments therein, it is likely that vicuña populations differ to some extent due to adaptation to local environmental variables. We propose the use of continuous MU for vicuña conservation and that this data serve as a baseline for genetic variation monitoring to avoid further loss of genetic diversity in captivity.

ETHICS STATEMENT

Samples were collected following guidelines of the American Society of Mammalogists (Sikes et al., 2011). Specific permits were required for the Servicio Agrícola y Ganadero, SAG (permit 447, 2002), the Corporación Nacional Forestal, CONAF (permit 6/02, 2002), for granting other collection permits and helping in collecting samples. The animal research oversight committee of Universidad del Bío-Bío had knowledge of sampling plans prior to their approval of the present animal research protocol. All experimental protocols were approved by the Institutional Animal Care and Use Committee of Universidad del Bío-Bío, the methods were carried out in accordance with the approved guidelines.

AUTHOR CONTRIBUTIONS

JM developed the ideas and obtained funding for the project. JW, BG, and JM collected the samples. JV, JC, RR, NA, and AC conducted the DNA analyses. AC, AA, DG-U, VV, and PO-tW analyzed the data. JM, BG, DG-U, JW, and PO-tW wrote the manuscript. All authors read, commented on and approved the final manuscript.

FUNDING

In Chile this research was supported by FONDECYT grant 1140785, CONICYT (Beca de Apoyo a Tesis Doctoral and Redes Internacionales para Investigadores en Etapa Inicial REDI170208). DG-U receives funding through Núcleo Milenio INVASAL from Iniciativa Científica Milenio, sponsored by Chile's Ministerio de Economía, Fomento y Turismo. In Peru, research was supported by Darwin Initiative for the Survival of Species (United Kingdom) grant 162/06/126 (1997–2000), The British Embassy (Lima), NERC (United Kingdom) grant GST/02/828 (1994–1998), the European Commission INCO-DC ICA4-2000-10229, MACS (2001–2005), and a Newton Fund Researcher Links Travel grant (ID: RLTG9-LATAM-359537872) funded by the United Kingdom Department for Business, Energy and Industrial Strategy and CONCYTEC (Peru) and delivered by the British Council; Asociación Ancash, Peru (2004); FINCyT Perú grant 006-FINCyT-PIBAP-2007 (2008–2010); and COLP, Compañía Operadora de LNG del Perú contract PLNG-EV-09012 (2010–2011).

ACKNOWLEDGMENTS

In Chile, we thank the Servicio Agrícola y Ganadero, SAG, the Corporación Nacional Forestal, CONAF for granting other collection permits and help in collecting samples Cristian Bonacic (Pontificia Universidad Católica de Chile), Pablo Valdecantos (Universidad Nacional de Tucumán), Bibiana Vilá (Universidad de Lujan), Luis Jacome (Zoológico de Buenos Aires, Argentina) and Alberto Duarte (Zoológico de Mendoza, Argentina) for sharing samples. Samples were collected under permits from CONACS, SERNANP and DGFFS (details above). Special thanks go to Carlos Loret de Mola and Maria Luisa del Rio (CONAM), Domingo Hoces, Wilder Trejo, Daniel Rivera, Daniel Arestegui, Leonidas Rodriguez, and Dirky Arias (CONACS); Gustavo Suarez de Freitas and Antonio Morizaki (INRENA); Felipe San Martin (Facultad de Medicina Veterinaria, Universidad Nacional Mayor de San Marcos) and Alejandro Camino (Asociacion Ancash). Over the years, David Perez, Antony Rodriguez, Juan Manuel Aguilar, and Paloma Krüger have helped in collecting samples. At CONOPA, Domingo Hoces, Lenin Maturrano, Alvaro Veliz, Katherine Yaya, Juan Olazabal, Raul Rosadio, and Michael Bruford have all contributed to the research reported here – from sampling to laboratory analysis and interpretation of the results.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00445/full#supplementary-material

REFERENCES

Acebes, P., Wheeler, J., Baldo, J., Tuppia, P., Lichtenstein, G., Hoces, D., et al. (2018). *Vicugna vicugna. The IUCN Red List of Threatened Species 2018: e.T22956A18540534*. Available at: https://www.iucnredlist.org/species/22956/18540534 (accessed May 4, 2019).

Aldenderfer, M. (1999). The Pleistocene/Holocene transition in Peru and its effects upon human use of the landscape. *Quat. Int.* 53/54, 11–19. doi: 10.1016/s1040-6182(98)00004-4

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-terWengel, P., Pitt, D., et al. (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6707–6712. doi: 10.1073/pnas.1519508113

Ammann, C., Jenny, B., Kammer, K., and Messerl, I. B. (2001). Late quaternary glacier response to humidity changes in the arid Andes of Chile (18–29° S). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 172, 313–326. doi: 10.1016/s0031-0182(01)00306-6

Anello, M., Daverio, M. S., Romero, S. R., Rigalt, F., Silbestro, M. B., and Vidal-Rioja, L. (2016). Genetic diversity and conservation status of managed vicuña (*Vicugna vicugna*) populations in Argentina. *Genetica* 144, 85–97. doi: 10.1007/s10709-015-9880-z

Arzamendia, Y., Carbajo, A. E., and Vilá, B. (2018). Social group dynamics and composition of managed wild vicuñas (*Vicugna vicugna vicugna*) in Jujuy, Argentina. *J. Ethol.* 36, 125–134. doi: 10.1007/s10164-018-0542-3

Baid, C. A., and Wheeler, J. (1993). Evolution of high Andean Puna ecosystems: environment, climate, and culture change over the last 12,000 years in the central Andes. *Mt. Res. Dev.* 13, 145–156.

Baigún, R. J., Bolkovic, M. L., Aued, M. B., Li Puma, M., and Scandalo, R. (2008). *Manejo de Fauna Silvestre en la Argentina. Primer Censo Nacional de Camélidos Silvestres al Norte del Río Colorado*. Buenos Aires: Secretaría de Ambiente y Desarrollo Sustentable de la Nación.

Banks, S. C., and Peakall, R. O. D. (2012). Genetic spatial autocorrelation can readily detect sex-biased dispersal. *Mol. Ecol.* 21, 2092–2105. doi: 10.1111/j.1365-294X.2012.05485.x

Blard, P. H., Sylvestre, F., Tripati, A. K., Claude, C., Causse, C., Coudrain, A., et al. (2011). Lake highstands on the Altiplano (Tropical Andes) contemporaneous with Heinrich 1 and the Younger Dryas: new insights from 14C, U–Th dating and δ18O of carbonates. *Quat. Sci. Rev.* 30, 3973–3989. doi: 10.1016/j.quascirev.2011.11.001

Boswall, J. (1972). Vicuna in Argentina. *Oryx* 11, 449–456.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Brohede, J., Primmer, C. R., Moller, A., and Ellegren, H. (2002). Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. *Nucleic Acids Res.* 30, 1997–2003. doi: 10.1093/nar/30.9.1997

Bulut, Z., McCormick, C. R., Gopurenko, D., Williams, R. N., Bos, D. H., and DeWoody, J. A. (2009). Microsatellite mutation rates in the eastern tiger salamander (*Ambystoma tigrinum tigrinum*) differ 10-fold across loci. *Genetica* 136, 501–504. doi: 10.1007/s10709-008-9341-z

Canedi, A. A., and Pasini, P. S. (1996). Repoblamiento y bioecologia de la vicuna silvestre en la Provincia de Jujuy, Argentina. *Anim. Genet. Resour.* 18, 7–21. doi: 10.1017/s1014233900000663

Casey, C. S., Orozco-terWengel, P., Yaya, K., Kadwell, M., Fernández, M., and Marín, J. C. (2018). Comparing genetic diversity and demographic history in codistributed wild South American camelids. *Heredity* 121, 387–400. doi: 10.1038/s41437-018-0120-z

Caughley, G. (1994). Directions in conservation biology. *J. Anim. Ecol.* 63, 215–244.

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., and Beaumont, M. A. (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* 186, 983–995. doi: 10.1534/genetics.110.118661

Ciofi, C., Beaumontf, M. A., Swingland, I. R., and Bruford, M. W. (1999). Genetic divergence and units for conservation in the Komodo dragon *Varanus komodoensis*. *Proc. R. Soc. Lond. B Biol. Sci.* 266, 2269–2274. doi: 10.1098/rspb.1999.0918

Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1659. doi: 10.1046/j.1365-294x.2000.01020.x

Courtois, R., Bernatchez, L., Ouellet, J.-P., and Breton, L. (2003). Significance of caribou (*Rangifer tarandus*) ecotypes from a molecular genetics viewpoint. *Conserv. Genet.* 4, 393–404. doi: 10.1023/A:1024033500799

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatikin, M., and Frimer, N. B. (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. U.S.A.* 91, 3166–3170. doi: 10.1073/pnas.91.8.3166

Escalante, M. A., García-De-León, F. J., Dillman, C. B., de los Santos Camarillo, A., George, A., and Barriga-Sosa, I. D. L. A. (2014). Genetic introgression of cultured rainbow trout in the Mexican native trout complex. *Conserv. Genet.* 15, 1063–1071. doi: 10.1007/s10592-014-0599-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294x.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.3410/f.1015548.197423

Franklin, W. L. (1982). Biology, ecology, and relationship to man of the South American camelids. *Mamm. Biol. South Am.* 6, 457–489.

Franklin, W. L. (1983). "Contrasting socioecologies of South America's wild camelids: the vicuña and the guanaco," in *Advances in the Study of Mammalian Behavior Special Publication Number 7*, eds J. F. Eisenberg and D. G. Kleinman (Provo, UT: The American Society of Mammalogists), 573–629.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147, 915–925.

González, B. A., Marín, J. C., Toledo, V., and Espinoza, E. (2016). Wildlife forensic science in the investigation of poaching of vicuña. *Oryx* 50, 14–15. doi: 10.1017/s0030605315001295

González, B. A., Palma, R. E., Zapata, B., and Marín, J. C. (2006). Taxonomic and biogeographical status of guanaco *Lama guanicoe* (Artiodactyla, Camelidae). *Mamm. Rev.* 36, 157–178. doi: 10.1111/j.1365-2907.2006.00084.x

Goudet, J. (2005). FSTAT (Version 2.9.4.: A.): *A Program to Estimate and Test Population Genetics Genetics Parameters*. Available at: http://www2.unil.ch/popgen/softwares/fstat.htm

Grimwood, I. (1969). *Notes on the Distribution and Status of Some Peruvian Mammals*, Vol. 21. New York, NY: Zoological Society.

Jost, L. (2008). GST and its relatives do not measure differentiation. *Mol. Ecol.* 17, 4015–4026. doi: 10.1111/j.1365-294x.2008.03887.x

Jungius, H. (1972). Bolivia and the vicuna. *Oryx* 11, 335–346.

Koford, C. B. (1957). The vicuña and the puna. *Ecol. Monogr.* 27, 153–219.

Kull, C., Grosjean, M., and Veit, H. (2002). Modeling modern and late Pleistocene glacio-climatological conditions in the north Chilean Andes (29–30). *Clim. Change* 52, 359–381. doi: 10.1023/A:1013746917257

Lang, K. D. M., Wang, Y., and Plante, Y. (1996). Fifteen polymorphic dinucleotide microsatellites in llamas and alpacas. *Anim. Genet.* 27, 293–293. doi: 10.1111/j.1365-2052.1996.tb00502.x

Marín, J. C. (2004). *Filogenia Molecular, Filogeografía y Domesticación de Camelidos Sudamericanos (ARTIODACTYLA: CAMELIDAE)*. Ph.D. thesis, Universidad de Chile, Santiago.

Marín, J. C., Casey, C. S., Kadwell, M., Yaya, K., Hoces, D., Olazabal, J., et al. (2007a). Mitochondrial phylogeography and demographic history of the vicuña: implications for conservation. *Heredity* 99, 70–80. doi: 10.1038/sj.hdy.6800966

Marín, J. C., Zapata, B., González, B. A., Bonacic, C., Wheeler, J. C., Casey, C., et al. (2007b). Systematics, taxonomy and domestication of alpaca and llama: new chromosomal and molecular evidence. *Rev. Chil. Hist. Nat.* 80, 121–140. doi: 10.4067/S0716-078X2007000200001

Marín, J. C., Romero, K., Rivera, R., Johnson, W. E., and González, B. A. (2017). Y-chromosome and mtDNA variation confirms independent domestications and directional hybridization in South American camelids. *Anim. Genet.* 48, 591–595. doi: 10.1111/age.12570

Marín, J. C., González, B. A., Poulin, E., Casey, C. S., and Johnson, W. E. (2013a). The influence of the arid Andean high plateau on the phylogeography and population genetics of guanaco (*Lama guanicoe*) in South America. *Mol. Ecol.* 22, 463–482. doi: 10.1111/mec.12111

Marín, J. C., Varas, V., Vila, A. R., López, R., Orozco-terWengel, P., and Corti, P. (2013b). Refugia in patagonian fjords and the eastern Andes during the last glacial maximum revealed by huemul (*Hippocamelus bisulcus*) phylogeographical patterns and genetic diversity. *J. Biogeogr.* 40, 2285–2298. doi: 10.1111/jbi.12161

Meirmans, P. G., and Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol. Ecol. Notes* 4, 792–794. doi: 10.1111/j.1471-8286.2004.00770.x

Moodley, Y., Russo, I. R. M., Dalton, D. L., Kotzé, A., Muya, S., Haubensak, P., et al. (2017). Extinctions, genetic erosion and conservation options for the black rhinoceros (*Diceros bicornis*). *Sci. Rep.* 7:41417. doi: 10.1038/srep41417

Moraes-Barros, N., Miyaki, C. Y., and Morgante, J. S. (2007). Identifying management units in non-endangered species: the example of the sloth *Bradypus variegatus* Schinz, 1825. *Braz. J. Biol.* 67, 829–837. doi: 10.1590/s1519-69842007000500005

Mosa, S. G. (1987). Tasa de natalidad de una población de vicuñas (Vicugna vicugna Miller, 1931) en estado de semicautividad. *An. Mus. Hist. Natural Valparaíso* 18, 177–179.

Napolitano, C., Díaz, D., Sanderson, J., Johnson, W. E., Ritland, K., Ritland, C. E., et al. (2015). Reduced genetic diversity and increased dispersal in guigna (*Leopardus guigna*) in Chilean fragmented landscapes. *J. Hered.* 106, 522–536. doi: 10.1093/jhered/esv025

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York, NY: Columbia University Press.

Nielsen, E. E., Bach, L. A., and Kotlicki, P. (2006). HYBRIDLAB (version 1.0): a program for generating simulated hybrids from population samples. *Mol. Ecol. Notes* 6, 971–973. doi: 10.1111/j.1471-8286.2006.01433.x

Park, S. D. E. (2001). *Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection*. Ph.D. thesis, University of Dublin, Dublin.

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Penedo, M. C. T., Caetano, A., and Cordova, K. I. (1998). Microsatellite markers for South American camelids. *Anim. Genet.* 29, 411–412. doi: 10.1046/j.1365-2052.1999.00526-21.x

Penedo, M. C. T., Caetano, A. R., and Cordova, K. (1999). Eight microsatellite markers for South American camelids. *Anim. Genet.* 30, 166–167. doi: 10.1046/j.1365-2052.1999.00382-8.x

Peter, B. M., Wegmann, D., and Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol. Ecol.* 19, 4648–4660. doi: 10.1111/j.1365-294X.2010.04783.x

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6, 7–11.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Pritchard, J. K., Wen, X., and Falush, D. (2007). *Documentation for Structure Software: version 2.2.* Chicago, IL: University of Chicago.

Rannala, B. (2007). *BayesAss Edition 3.0 User's Manual*.Available at: http://www.rannala.org/?page_id=245 (accessed November 27, 2012).

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rodriìguez, F., and Morales-Delanuez, A. (2017). *La Vicunþa Ecuatoriana y su Entorno*, 1st Edn. Santo Domingo: Ministerio del Ambiente de Ecuador.

Ruiz-Garcia, M., Orozco-terWengel, P., Castellano, A., and Arias, L. (2005). Microsatellite analysis of the spectacled bear (*Tremarctos ornatus*) across its range distribution. *Genes Genet. Syst.* 80, 57–69. doi: 10.1266/ggs.80.57

Sambrook, J., Fritschi, E. F., and Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. New York, NY: Cold Spring Harbor Laboratory Press.

Sarno, R. J., David, V. A., Franklin, W. L., O'Brien, S. J., and Johnson, W. E. (2000). Development of microsatellite markers in the guanaco, *Lama guanicoe*: utility for South American camelids. *Mol. Ecol.* 9, 1922–1924. doi: 10.1046/j.1365-294x.2000.01077-3.x

Sarno, R. J., Villalba, L., Bonacic, C., González, B., Zapata, B., Mac Donald, D. W., et al. (2004). Phylogeography and subspecies assessment of vicunas in Chile and Bolivia utilizing mtDNA and microsatellite markers: implications for vicuna conservation and management. *Conserv. Genet.* 5, 89–102. doi: 10.1023/b:coge.0000014014.01531.b6

Schneider, S., and Excoffier, L. (1999). Estimation of demographic para- meters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* 152, 1079–1089.

Sikes, R. S., Gannon, W. L., and The Animal Care and use Committee of the American Society of Mammalogists (2011). Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *J. Mammal.* 92, 235–253. doi: 10.1093/jmammal/gyw078

Stølen, K. A., Lichtenstein, G., and Nadine, R. D. A. (2009). "Local participation in vicuña management," in *The Vicuña*, ed. I. J. Gordon (Boston, MA: Springer), 81–96.

Storz, J. F., and Beaumont, M. (2002). Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical bayesian model. *Evolution* 56, 154–166. doi: 10.1111/j.0014-3820.2002.tb00857.x

Sveegaard, S., Galatius, A., Dietz, R., Kyhn, L., Koblitz, J. C., Amundin, M., et al. (2015). Defining management units for cetaceans by combining genetics, morphology, acoustics and satellite tracking. *Glob. Ecol. Conserv.* 3, 839–850. doi: 10.1016/j.gecco.2015.04.002

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Taylor, B. L., and Dizon, A. E. (1999). First policy then science: why a management unit based solely on genetic criteria cannot work. *Mol. Ecol.* 8, S11–S16.

Van Oosterhout, C., Hutchinson, W., Willis, D., and Shipley, P. (2004). MICROCHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* 4, 535–538. doi: 10.1111/j.1471-8286.2004.00684.x

Vicuña Convention (2017). *33rd Ordinary Meeting of the Vicuɲa Convention*. Jujuy, Argentina.

Villalba, L., Cuéllar, E., and Tarifa, T. (2010). "Capítulo 23 camelidae," in *Distribución, Ecología y Conservación de los Mamíferos Medianos y Grandes de Bolivia*, eds R. B. Wallace, H. Gómez, Z. R. Porcel, and D. I. Rumiz (Santa Cruz de la Sierra: Centro de Ecología y Difusión Simón I. Patino).

Wallace, R. B., Gómez, H., Porcel, Z. R., and Rumiz, D. I. (2010). *Distribución, Ecología y Conservación de los Mamíferos Medianos y Grandes de Bolivia*. Santa Cruz de la Sierra: Centro de Ecología y Difusión Simón I. Patino.

Weber, J. L., and Wong, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2, 1123–1128. doi: 10.1093/hmg/2.8.1123

Weeks, A. R., Stoklosa, J., and Hoffmann, A. A. (2016). Conservation of genetic uniqueness of populations may increase extinction likelihood of endangered species: the case of Australian mammals. *Front. Zool.* 13:31. doi: 10.1186/s12983-016-0163-z

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-Statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x

Wheeler, J., Pires-Ferreira, E., and Kaulicke, P. (1976). Preceramic animal utilization in the Central Peruvian Andes. *Science* 194, 483–490. doi: 10.1126/science.194.4264.483

Wheeler, J. C. (1995). Evolution and present situation of the South American Camelidae. *Biol. J. Linn. Soc.* 54, 271–295. doi: 10.1111/j.1095-8312.1995.tb01037.x

Wheeler, J. C. (2006). "Historia natural de la vicuña," in *Investigacion Conservacion y Manejo de Vicunas*, ed. B. Vilá (Buenos Aires: Proyecto MACS), 25–36.

Wheeler, J. C. (2012). South American camelids - past, present and future. *J. Camelid Sci.* 5, 1–24. doi: 10.3389/fpls.2018.00649

Wheeler, J. C., Fernández, M., Rosadio, R., Hoces, D., Kadwell, M., and Bruford, M. W. (2001). Diversidad genética y manejo de poblaciones de vicuñas en el Perú. *RIVEP Rev. Investig. Vet. Perú* 1, 170–183.

Wheeler, J. C., and Laker, J. (2009). "The vicuña in the Andean altiplano," in *The Vicuña: the theory and Practice of Community-Based Wildlife Management*, ed. I. Gordon (Boston, MA: Springer), 21–33. doi: 10.1007/978-0-387-09476-2_3

Wilson, G. A., and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163, 1177–1191.

Yacobaccio, H. (2009). "The historical relationship between people and the vicuña," in *The Vicuña*, ed. I. J. Gordon (Boston, MA: Springer), 7–20. doi: 10.1007/978-0-387-09476-2_2

Yannic, G., St-Laurent, M. H., Ortego, J., Taillon, J., Beauchemin, A., Bernatchez, L., et al. (2016). Integrating ecological and genetic structure to define management units for caribou in Eastern Canada. *Conserv. Genet.* 17, 437–453. doi: 10.1007/s10592-015-0795-0

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 González, Vásquez, Gómez-Uchida, Cortés, Rivera, Aravena, Chero, Agapito, Varas, Wheleer, Orozco-terWengel and Marín. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 June 2019
doi: 10.3389/fgene.2019.00502

Beyond the Big Five: Investigating Myostatin Structure, Polymorphism and Expression in *Camelus dromedarius*

Maria Favia[1], Robert Fitak[2,3], Lorenzo Guerra[1], Ciro Leonardo Pierri[1], Bernard Faye[4], Ahmad Oulmouden[5], Pamela Anna Burger[2] and Elena Ciani[1]*

[1] Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari "Aldo Moro", Bari, Italy, [2] Research Institute of Wildlife Ecology, Vetmeduni, Vienna, Austria, [3] Department of Biology, Duke University, Durham, NC, United States, [4] CIRAD, UMR SELMET, Montpellier, France, [5] Département Sciences du Vivant, Université de Limoges, Limoges, France

Myostatin, a negative regulator of skeletal muscle mass in animals, has been shown to play a role in determining muscular hypertrophy in several livestock species, and a high degree of polymorphism has been previously reported for this gene in humans and cattle. In this study, we provide a characterization of the myostatin gene in the dromedary (*Camelus dromedarius*) at the genomic, transcript and protein level. The gene was found to share high structural and sequence similarity with other mammals, notably Old World camelids. 3D modeling highlighted several non-conservative SNP variants compared to the bovine, as well as putative functional variants involved in the stability of the myostatin dimer. NGS data for nine dromedaries from various countries revealed 66 novel SNPs, all of them falling either upstream or downstream the coding region. The analysis also confirmed the presence of three previously described SNPs in intron 1, predicted here to alter both splicing and transcription factor binding sites (TFBS), thus possibly impacting myostatin processing and/or regulation. Several putative TFBS were identified in the myostatin upstream region, some of them belonging to the myogenic regulatory factor family. Patterns of SNP distribution across countries, as suggested by Bayesian clustering of the nine dromedaries using the 69 SNPs, pointed to weak geographic differentiation, in line with known recurrent gene flow at ancient trading centers along caravan routes. Myostatin expression was investigated in a set of 8 skeletal muscles, both at transcript and protein level, via Digital Droplet PCR and Western Blotting, respectively. No significant differences were observed at the transcript level, while, at the protein level, the only significant differences concerned the promyostatin dimer (75 kDa), in four pair-wise comparisons, all involving the *tensor fasciae latae* muscle. Beside the mentioned band at 75 kDa, additional bands were observed at around 40 and 25 kDa, corresponding to the promyostatin monomer and the active C-terminal myostatin dimer, respectively. Their weaker intensity suggests that the unprocessed myostatin dimers could act as important reservoirs of slowly available myostatin forms. Under this assumption, the sequential cleavage steps may contribute additional layers of control within an already complex regulatory framework.

Keywords: *Camelus dromedarius*, myostatin, skeletal muscle, Single Nucleotide Polymorphisms, Next Generation Sequencing, Digital Droplet PCR, Western Blot, 3D protein comparative modeling

OPEN ACCESS

Edited by:
Edward Hollox, University of Leicester, United Kingdom

Reviewed by:
Kieran G. Meade, Teagasc, The Irish Agriculture and Food Development Authority, Ireland
René Massimiliano Marsano, University of Bari Aldo Moro, Italy

****Correspondence:***
Lorenzo Guerra
lorenzo.guerra1@uniba.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 January 2019*
Accepted: *07 May 2019*
Published: *07 June 2019*

Citation:
Favia M, Fitak R, Guerra L, Pierri CL, Faye B, Oulmouden A, Burger PA and Ciani E (2019) Beyond the Big Five: Investigating Myostatin Structure, Polymorphism and Expression in Camelus dromedarius. Front. Genet. 10:502. doi: 10.3389/fgene.2019.00502

INTRODUCTION

Myostatin (alias growth and differentiation factor-8, GDF8), a member of the transforming growth factor-β (TGF-β) super-family, is a negative regulator of skeletal muscle mass in animals during development and growth. It is exclusively expressed in skeletal muscle during embryogenesis, while in adults is also detected, at a much lower level, in other tissues (e.g., heart, adipose tissue, mammary gland) (McPherron et al., 1997; Ji et al., 1998; Sharma et al., 1999; Morissette et al., 2006; Shyu et al., 2006; Allen et al., 2008). Expression in these tissues can be upregulated under pathological conditions, such as myocardial infarction (Sharma et al., 1999), obesity (Allen et al., 2008) or experimentally induced skeletal muscle atrophy (Rodriguez et al., 2014), while it can be down regulated during chronic exercise (Carlson et al., 1999; Reardon et al., 2001; Kim et al., 2005; Matsakas et al., 2006; Allen et al., 2009; Gustafsson et al., 2010).

Like other TGF-β super-family members, myostatin is synthesized as a precursor protein (375 amino acids), referred to as pre-promyostatin. After translocation to the endoplasmic reticulum, it goes through a first cleavage to remove a 24-amino acid signal peptide and it forms a disulfide-linked homodimer (promyostatin dimer). Within the Golgi, the promyostatin dimer may be further cleaved by the furin family of protein convertases to generate two NH_2-terminal (27.7 kDa, each) and two disulfide-linked COOH-terminal fragments (12.4 kDa, each) (Lee and McPherron, 2001; Thies et al., 2001). The two NH_2-terminal fragments (also referred to as pro-domains) may complex with the COOH-terminal dimer (also referred to as active myostatin) via a non-covalent bound that maintains myostatin in a latent state by rendering it unable to engage its receptors (Wolfman et al., 2003; Jiang et al., 2004). The "latent myostatin complex" (Lee and McPherron, 2001; Thies et al., 2001) may be secreted in the extracellular space, though it has been shown that, in skeletal myocytes, myostatin is mainly secreted as uncleaved promyostatin (Anderson et al., 2008; Pirruccello-Straub et al., 2018). In the extracellular space, the action of furin protein convertase and metalloproteinases (like BMP-1, TLL-1, and TLL-2) may finally convert the uncleaved promyostatin and the latent complex, respectively, into the active form of myostatin (Lakshman et al., 2009). Notwithstanding, in serum, myostatin has been shown to exist mainly as a latent complex (Hill et al., 2002; Zimmers et al., 2002; Lee, 2010). The active myostatin present in plasma circulates bound to several proteins (Miura et al., 2006; Cash et al., 2009; Walker et al., 2015), including follistatin, FSTL3, GASP1, GASP2 and decorin, that prevent it binding to the receptor and activating signaling (Hill et al., 2002, 2003; Lee, 2008; Cotton et al., 2018). Composite pools of myostatin are hence available at the various compartments, suggesting that extracellular processing of the protein may be a key regulatory step for its signaling (Anderson et al., 2008). The presence, at various extents, of the myostatin active form in the extracellular space and in the serum is consistent with the postulated autocrine, paracrine (Gao et al., 2013), and/or endocrine manner of function (Zimmers et al., 2002).

Upon binding to the target cell, myostatin induces the formation of a heterotetrameric complex made of two activin responsive type II receptors (ActIIRA or, preferentially, ActIIRB) and two type I receptors, either activin (ALK4) or TGF-β (ALK5) (Lee and McPherron, 2001; Rebbapragada et al., 2003). Signaling is hence initiated by phosphorylation of SMAD2 or SMAD3, operated by the type I receptors, followed by translocation of SMADs to the nucleus for modulation of gene expression (Huang et al., 2011). In particular, in skeletal muscle, myostatin is known to block the transcription of genes responsible for the myogenesis, among which MyoD, a transcriptional factor that is involved in skeletal muscle development and repair (Megeney et al., 1996; Cornelison et al., 2000; Guttridge et al., 2000; Montarras et al., 2000). Beside the above mentioned canonical pathway, two other pathways have been highlighted, involving MAPK activation or inhibition of Akt signaling (Elkina et al., 2011).

Myostatin genomic organization was first provided for the murine species by McPherron et al. (1997) who also reported on the highly conserved nature of the myostatin transcript across several species. The myostatin gene (*MSTN*) comprises three exons and two introns. The nucleotide sequence coding for the active form of myostatin (109 a.a) is located in the 3′ terminal of the third exon (Gonzalez-Cadavid et al., 1998). Effects of abolishing myostatin function were first explored by McPherron et al. (1997) in mutant mice where the entire mature C-terminal region was deleted, showing a two- to three-fold increase in skeletal muscle mass in mutant compared to wild-type animals. Mutations at the myostatin gene, responsible for a significantly increased skeletal muscle mass, were also shown to naturally occur in several livestock species, like cattle, sheep, pigs, dogs, horses, rabbit, poultry (for a review, see Aiello et al., 2018), and human (Schuelke et al., 2004). In particular, the high level of polymorphism previously described for the myostatin gene in humans (Ferrell et al., 1999) was confirmed in cattle in a survey of 678 animals from 28 European breeds by using Single Strand Conformation Polymorphism (SSCP) analysis, followed by Sanger sequencing of the PCR re-amplified SSCP bands (Dunner et al., 2003). A total of 10 silent, 3 missense and 6 disruptive mutations were detected in the above study, giving origin to 20 distinct haplotypes whose sequence variation and breed distribution patterns supported the hypothesis that origin of muscular hypertrophy (also known in cattle as "double muscle" phenotype) was the result of both (i) European dispersal of the common variant nt821(del11) and (ii) arising and maintaining of various (mostly disruptive) mutations in single breeds.

Old World camels include both wild (*Camelus ferus*) and domestic (*Camelus dromedarius* and *Camelus bactrianus*) species. Despite differences in muscularity can be observed among distinct populations and/or individuals, these are not dramatic as those observed in other livestock species and no evident "double muscle" phenotype has been described so far (B. Faye, personal communication). The myostatin gene has been previously characterized in various dromedary populations from Pakistan and India, although only 256 bp of exon 1 and 375 bp of exon 2, respectively, were considered in the analyses (Shah et al., 2006; Agrawal et al., 2017). A more comprehensive sequence polymorphism analysis of the myostatin gene was performed in our laboratories, where more than 3.6 kb of genomic sequence,

including the three exons, small part of the introns and part of the 3′ and 5′ ends, was sequenced in a total of 22 dromedaries from three different Northern African geographic regions (Muzzachi et al., 2015). In this study, to further expand the knowledge base about myostatin, we followed up by (i) characterizing the gene structure (transcriptional initiation/termination sites; exon/intron boundaries), (ii) analyzing polymorphism of the complete genomic sequence and of the partial cDNA in a set of animals from various sampling sites, (iii) investigating expression patterns at both the transcript and the protein level in different skeletal muscles.

MATERIALS AND METHODS

Characterization of the Full-Length Myostatin cDNA

RNA Isolation From Skeletal Muscles

Skeletal muscles were sampled at slaughterhouses from seven different animals (Sudan, 3; Egypt, 2; Mauritania, 2). For each animal, a small sample of frozen muscle tissue (100 mg), previously stored in tubes containing RNAlater (QIAGEN), was finely chopped by using a sharp scalpel in 2 ml RLT buffer (RNeasy Midi Kit, QIAGEN) supplemented with β-mercaptoethanol following the manufacturer's instructions. The sample was then homogenized using the T 10 basic Ultra-Turrax homogenizer (IKA). After homogenization, the sample was added with 4 ml of RNase-/DNase-free water plus 65 μl of Proteinase K (SIGMA, ≥0.6 Units/μl). After an incubation step at 55°C for 20 min, samples were processed following manufacturer's instruction.

Myostatin Transcription Initiation and Termination Sites

Transcription initiation and termination sites were identified using the RACE (Rapid amplification of cDNA ends) PCR approach implemented in the SMARTer RACE cDNA Amplification Kit (CloneTech) following manufacturer's instructions.

For the 5′RACE-PCR, the following primers were used:

- 1st reaction: 5′ATCCTCAGTAAACTTCGCCTGGAAACAGCT3′
- 2nd reaction: 5′GGCTGTGTAATGCATGTATGTGGAGACAAA3′

For the 3′RACE-PCR, the following primers were used:

- 1st reaction: 5′TGTGCACCAAGCAAACCCCAGAGGTTCGGC3′
- 2nd reaction: 5′CCTGCTGTACTCCCACAAAGATGTCTCCAA3′

After separation on a 2% agarose gel, PCR products were excised from the gel, purified using the QIAquick Gel Extraction Kit (QIAGEN) and sequenced using the following primers:

5′RACE: 5′TTTGTCTCCACATACATGCATTACACAGCC3′

3′RACE: 5′CCTGCTGTACTCCCACAAAGATGTCTCCAA3′

RNA Retro-Transcription

After quality control using a NanoDrop 2000C spectrophotometer (Thermo Fisher Scientific), 50 μl of total RNA was retro-transcribed into cDNA using High Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific), which is based on a combination of oligo(dT) and random primers, following manufacturer's instruction.

Amplification and Sequencing of the Myostatin cDNA

A nested PCR was developed, with external forward and reverse primers falling in the 5′UTR and 3′UTR, respectively. The primer sequences were:

- Forward 5′CCTTGGCATTACTCAAAAGCAA3′
- Reverse 5′CCTAAGTTTTCGAGCTAGGAGATC3′

The PCR conditions were: initial step at 95°C for 2 min; 35 cycles of a three-step thermal profile of 95°C for 30 s (denaturation), 57°C for 30 s (annealing), 72°C for 60 s (elongation); a final elongation step at 72°C for 5 min.

Internal primers were:

- Forward 5′CAGTACGATGTCCAGAGAGATGACAGCAGT3′
- Reverse 5′TGTGCACCAAGCAAACCCCAGAGGTTCGGC3′

The PCR conditions were: initial step at 95°C for 2 min; 35 cycles of a three-step thermal profile of 95°C for 30 s (denaturation), 61°C for 30 s (annealing), 72°C for 60 s (elongation); a final elongation step at 72°C for 5 min.

For both reactions, 3 μl of cDNA were added to a solution of 12.5 μl Master Mix (Multiplex PCR Kit, QIAGEN), 1 μl of each primer (Forward and Reverse), 7.5 μl water. After separation on a 2% agarose gel, PCR products were excised from the gel, purified using the QIAquick Gel Extraction Kit (QIAGEN) and sequenced on both directions with internal primers, using the Sanger method.

Sequences Alignment

Sequences obtained via Sanger method (RACE PCR and cDNA amplicons, see above) were aligned using ClustalOmega (Sievers et al., 2011).

Precursor Prediction

The SignalP 4.1[1], the Combined Signal Peptide Predictor (CoSiDe)[2] and the Signal-3L 2.0[3] online tools were interrogated for predicting the most probable location of the signal peptide cleavage site.

Comparative Protein Modeling

Myostatin orthologs were searched through and sampled from *Mammalia*. The crystallized structure of the myostatin was available under the PDB_ID 5ntu (Cotton et al., 2018). The retrieved sequences, including the proposed crystallized structure

[1]http://www.cbs.dtu.dk/services/SignalP/

[2]http://sigpep.services.came.sbg.ac.at/coside.html

[3]http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/

(5ntu.pdb), were thus aligned by using ClustalW (see Pierri et al., 2010, and references therein) for investigating chemical-physical properties of the amino acid regions showing variants between the human and the dromedary sequences. Accession numbers of the sequences considered in this study, together with additional details concerning the databases and the online tools used in this study, are summarized in **Supplementary Table S1**.

Then, SPDBV was used for generating a 3D model of the dromedary myostatin protein according to protocols described in Pierri et al. (2010). The obtained 3D comparative model was energetically minimized. A total of 100 steps of energy minimization were performed for relaxing the obtained 3D model by using the energy minimization tools implemented in Chimera. WhatIF and Chimera biochemical tools were used for checking the correct 3D model packing. PyMOL was used for manual inspection of the investigated 3D models and for generating figures (see Pierri et al., 2010, and references therein).

Phylogenetic Analysis

The analysis of the evolutionary relationships among orthologous myostatin sequences was conducted using MEGA5 (Tamura et al., 2011). Orthologous sequences of myostatin/growth differentiation factor 8 with *E*-value lower than 10ˆ-55, query coverage higher than 70% and % of identical amino acids ranging between 40 and 100% were aligned by using ClustalW implemented in Jalview. For Arthropoda, Aves, and Mammalia, due to the existence of more than one hundred of sequences complying with the above criteria, we imposed a filter on the first 30 sequences for each taxonomic group. Redundant sequences with 100% identical amino acids were removed from the multiple sequence alignment. A final set of 83 protein sequences (see **Supplementary Data Sheet S1**) were retained for tree building. In detail, the tree was built from the ungapped multiple sequence alignment applying the maximum likelihood method with the JTT model for the amino acid substitutions and a gamma distribution (five discrete gamma categories) for the rates among sites. A total of 100 bootstrap samplings were applied to test the robustness of the tree.

Polymorphism Analysis

Sample Collection and Whole-Genome Sequencing

Whole blood from 25 Old World camels was collected during routine veterinary procedures or as part of a monitoring program of the wild camel population in Mongolia. These samples included nine dromedaries (*C. dromedarius*), seven domestic Bactrian camels (*C. bactrianus*), and nine wild camels (*C. ferus*). Dromedary camels were selected to represent a variety of geographic locations: Pakistan (1), Kenya (1), Kingdom of Saudi Arabia (3), Sudan (1), United Arab Emirates (1), Qatar (1), Canary Islands – Spain (1). Domestic Bactrian camel originated from Mongolia or Kazakhstan, while all the wild camels originated from Mongolia. DNA was extracted using the Master PureTM DNA purification kit for blood (Epicentre version III) and generated a 500 bp paired-end library for each sample. Each library was sequenced with a single lane of an Illumina HiSeq (Illumina, United States) according to standard protocols.

Whole-Genome Read Processing and Alignments

The 3′ end of sequence reads were trimmed to a minimum phred-scaled base quality score of 20 (probability of error $< 1.0\%$) and trimmed reads < 50 bp in length were excluded using POPOOLATION v1.2.2 (Kofler et al., 2011). All processed reads were aligned to the *C. ferus* CB1 reference genome (Genbank accession: AGVR01040332.1) using BWA v0.6.2 (Li and Durbin, 2009) with parameters '−n 0.01 −o 1 −e 12 −d 12 −l 32.' Duplicate reads were removed and alignments were filtered to only include reads that were properly paired and unambiguously mapped with a mapping quality score > 20. Reads around insertions/deletions were realigned and a base quality score recalibration was performed using the Genome Analysis Toolkit (GATK) v3.1-1 following guidelines presented by Van der Auwera et al. (2013). As input into the base quality score recalibration step, a stringently filtered set of single nucleotide variants (SNVs) was generated using the overlap of three different variant-calling algorithms [SAMTOOLS v1.1] (Li et al., 2009); [GATK HAPLOTYPECALLER v3.1-1] (Van der Auwera et al., 2013); [ANGSD v0.563] (Korneliussen et al., 2014). The overlapping SNVs were filtered to exclude those with a quality score (Q) < 20, depth of coverage (DP) > 750 (∼30X/individual), quality by depth (QD) < 2.0, strand bias (FS) > 60.0, mapping quality (MQ) < 40.0, inbreeding coefficient < -0.8, mapping quality rank sum test (MQRankSum) < -12.5, and read position bias (ReadPosRankSum) < -8.0. Furthermore, SNVs were excluded if three or more were found within a 20 bp window, were within 10 bp of an insertion/deletion, or were found in an annotated repetitive region.

Whole-Genome Variant Identification

Another set of SNVs from the realigned and recalibrated alignment files was generated using the GATK HAPLOTYPECALLER and filtering criteria as described above. SNVs on scaffolds putatively assigned to the X and Y chromosome, with a minimum allele count < 2, missing a genotype in more than five individuals, with $4 > DP > 30$ per genotype, and deviating from Hardy–Weinberg equilibrium ($p < 0.0001$) in VCFTOOLS v0.1.12b (Danecek et al., 2011) were further excluded. This set of SNVs was used as a training set to perform variant quality score recalibration in GATK, assigning a probability of error to the training set of 0.1. This recalibration develops a Gaussian mixture model across the various annotations in the high-quality training dataset then applies the model to all variants in the initial dataset. The process has been shown to outperform the 'hard' filtering of variants (e.g., Pirooznia et al., 2014). After variant recalibration, all SNVs with LOD score < -5.0 and $4 > DP > 30$ per genotype were excluded.

Identification and Characterization of Variants at the Myostatin Locus

The publicly available *Camelus ferus* myostatin sequence (GenBank Accession No AGVR01040332) was BLASTed against our Old World camel genomes and the contig-8645394 (*Camelus dromedarius*), contig-8938518 (*Camelus bactrianus*) and contig-7907533 (*Camelus ferus*) were retrieved and used in the comparative analysis of the myostatin locus at the

nucleotide level. Moreover, from the final set of SNVs described in the sub-section above, the *Camelus dromedarius* Single Nucleotide Polymorphisms (SNPs) falling in the contig-8645394 (**Supplementary Data Sheet S2**) were selected for further inspection.

Bayesian Clustering

The identified SNPs were used for clustering the nine dromedary samples by adopting the Bayesian algorithm implemented in the STRUCTURE software v. 2.2 (Falush et al., 2007). The analysis was performed without providing *a priori* information on population membership, adopting the "admixture model" option and a burn-in period of 10,000 generations, followed by 100,000 iterations. Five independent runs were performed for each *K* value (number of clusters to be tested), and the results were visually inspected for reproducibility. *K* values ranging from 1 to 9 were tested, and the *K* value showing the highest probability was discussed.

In silico Functional Prediction

The web-based analysis tool by the Human Splicing Finder Version 3.0.2 (Desmet et al., 2009), available at http://www.umd.be/HSF3/index.html, was used to predict putative functional effect of SNP variants in terms of potential alteration of splicing patterns. The *in silico* tool TFBIND (Tsunoda and Takagi, 1999), available at http://tfbind.hgc.jp/, was used to identify transcription factor binding sites (TFBS) and their possible disruption due to the presence of Single Nucleotide Polymorphisms.

Absolute Quantification of Myostatin Transcripts

RNA Isolation and cDNA Synthesis

Skeletal muscles were sampled at slaughterhouses (Kingdom of Saudi Arabia) from two adult animals. For each animal, eight muscles, representative of the different anatomical regions of the body were taken: *brachiocephalicus* (head/neck), *deltoid, extensor carpi radialis,* and *tensor fasciae latae* (forelimbs), *semitendinosus* and *coccygeus* (trunk), *biceps femoris,* and *peroneus longus* (hindlimbs). For all muscles, sampling occurred within 30 min *post-mortem*. Immediately after collection, samples were stored in tubes containing RNAlater (QIAGEN). For RNA isolation and cDNA synthesis, the procedures described above were adopted.

Digital Droplet PCR Assay Design

The Digital Droplet PCR method is based on end-point fluorescence signal detection, and the intensity of signal observed for positive droplets, varying with primer/template combination, is not considered for target quantification. However, in this system, droplets are interpreted as either "positive" or "negative," depending whether target amplification occurred or not, based on a settled fluorescence cut-off. Two different assays, with probes targeting the two possible exon junctions in the myostatin gene (between exon 1 and 2, and between exon 2 and 3), were used. FAM- (6-carboxy-fluorescein) and HEX- (hexa-chloro-fluorescein) labeled probes were used, respectively, in Assay 1 and Assay 2. Probes and primers sequences were as follows:

- Assay 1:
  Probe 1 5′-/56-FAM/CTACAGAGT/ZEN/CTGATCTTCT AATGC/3IABkFQ/-3′
  Primer 1 5′-GACGGAAACAATCATTACC-3′
  Primer 2 5′-GAGCTAAACTTAAAGAAGCAA-3′
- Assay 2:
  Probe 1 5′-/5HEX/AAGGGATTC/ZEN/AAACCATCTT CTC/3IABkFQ/-3′
  Primer 1 5′-GGTCATGATCTTGCTGTA-3′
  Primer 2 5′-GTCTGTTACCTTGACTTCTA-3′

By partitioning the reaction volume into thousands of droplets, this technique allows absolute quantification of nucleic acids without the need of a standard curve, with improved precision over classical quantitative PCR (qPCR).

Digital Droplet PCR Conditions

A 20-μl reaction mixture was prepared comprising of 10 μl ddPCR Supermix^{TM} for probes (no dUTP) (Bio-Rad), 1 μl primers and probe mix for Assay 1, 1 μl primers and probe mix for Assay 2, 2 μl cDNA, 6 μl RNase-/DNase-free water. The final concentration of primers and probe was 900 and 250 nM, respectively. The amplification conditions were 10 min DNA polymerase activation at 95°C, followed by 40 cycles of a two-step thermal profile of 30 s at 94°C for denaturation, and 60 s at 60°C for annealing and extension, followed by a final hold of 10 min at 98°C for droplet stabilization, and cooling to 4°C. A thermal cycler (T100^{TM}; Bio-Rad) was used, and the temperature ramp rate was set to 2°C/s, with the lid heated to 105°C, according to the Bio-Rad recommendations. A negative (no template) and a positive control were included. The latter consisted, for both assays, of a synthetic oligonucleotide (gBlocks Gene Fragment, by IDT), with a size of 467 bp, including junctions between exons 1 and 2 and between exons 2 and 3, designed based on the predicted sequence for the myostatin transcript in *Camelus dromedarius* (XM_010991955). In the reaction preparation, for the positive control, 2 μl of the above synthetic oligonucleotide were added, at a final concentration of 1 ng/ml. For all the considered muscles, two biological and two technical replicates were included in the experiment.

Data Analysis

After the thermal cycling, the plates were transferred to a droplet reader (QX200^{TM}; Bio-Rad). The software package provided with the ddPCR system was used for data acquisition (QuantaSoft^{TM} 1.6.6.0320; Bio-Rad). The rejection criterium for the exclusion of a reaction from subsequent analysis was a low number of droplets measured (<10,000 per 20 μl PCR). The data from the ddPCR are given in target copies/μl reaction. The significance of differences among muscles was tested using ANOVA (Analysis of Variance).

Protein Extraction and Western Blotting

A small sample of frozen muscle tissue (100 mg), previously stored in tubes containing RNAlater (QIAGEN), was finely chopped and homogenized by using a sharp scalpel in 300 μl Ripa buffer [10 mM Tris-HCl pH 7.4, 140 mM NaCl, 1% (v/v) Triton X-100, 1% (w/v) Na-deoxycholate, 0.1% (v/v) SDS, 1 mM

NaF, 1 mM EDTA, 1 mM Na_3VO_4] supplemented with 1x protease inhibitor cocktail (Sigma), and then by using the T 10 basic Ultra-Turrax homogenizer (IKA). After homogenization, the sample was kept on ice for 30 min, and then vortexed for 5 min. At the end, sample was centrifuged for 20 min at 4°C at 13,000 × g to remove unbroken cells, nuclei and cell debris. The supernatant, containing solubilized proteins, was recovered and protein concentration was measured by the method of Bradford (Bradford, 1976). An aliquot of 20 μg of protein for each sample was diluted in Laemmli buffer not containing DTT or β-mercaptoethanol, heated at 95°C for 5 min, and separated by 12 % (v/v) Tris/HCl SDS/PAGE. The separated proteins were transferred to Immobilon P (Millipore) in Trans-Blot semidry electrophoretic transfer cell (Amersham Biosciences) for immunoblotting. The used primary antibody was a rabbit polyclonal anti-MSTN antibody against the C-terminal region (300–349 aa) of mouse myostatin (TA343358, OriGene; dilution 1:1000) that presented broad species reactivity, including artiodactyls. The densitometric quantification and image processing of the considered bands were carried out using Adobe Photoshop and the Image software package (version 1.61, National Institutes of Health, Bethesda, MD, United States). The total lane density of transferred proteins on the membrane stained with Coomassie Blue dye was used for the normalization of the proteins of our interest. The significance of differences among muscles, for each considered band, was tested using ANOVA (Analysis of Variance). *Post hoc t*-tests were performed to determine where the groups differed. All *p*-levels for *post hoc t*-tests were adjusted using Bonferroni correction.

RESULTS

Myostatin Gene Organization

The RACE-PCR approach allowed to map the start transcription site (**Supplementary Figure S1A**) at 109 bp upstream the start-codon, in a position that is 24 and 25 bp downstream compared to the usual human and mouse transcription initiation sites, respectively[4]. The transcriptional termination site (**Supplementary Figure S1B**) was mapped 215 bp downstream the stop-codon, much earlier than in human and mouse where a 1561 and 1448 bp 3′UTR is usually reported[4]. No evidence was found, by combined RT-PCR and RACE approaches, for alternative splicing events or alternative 5′ or 3′ ends (**Supplementary Figure S2**). Based on the above, a 5292 bp genomic locus was identified for myostatin in *C. dromedarius*. The locus was highly conserved among the three Old World camelids species (**Supplementary Figure S3**). Comparative analysis of the *C. dromedarius* genomic sequence (contig-8645394) with the obtained cDNA sequences confirmed, as in other species, the presence of three exons and two introns, with a predicted *C. dromedarius* myostatin full length cDNA of 1452 bp (**Supplementary Figure S4A**) and a protein of 375 amino acids (**Figure 1**). The latter is consistent with the predicted protein size for most of the species[5]. The dromedary myostatin protein also showed all the hallmarks present in other TGF-ß family members, including an N-terminal signal sequence for secretion, a pro-region followed by the proteolytic processing RSRR site, and a C-terminal domain containing nine cysteine residues. In particular, the signal peptide was consistently predicted to have an 18 amino acid length by SignalP 4.1 and Signal-3L 2.0, while a length of 23 amino acids was predicted by CoSiDe.

Comparative 3D Protein Modeling

Sequences from nine *Artiodactyla* species, including the three phylogenetically close Old World camelids (*C. dromedarius, C. bactrianus,* and *C. ferus*), one New World camelid (*Vicugna pacos*), the two *Bos taurus* subspecies, i.e., *B. taurus taurus* (a non "double muscle" Hereford subject) and *B. taurus indicus*, the wild yak (*Bos mutus*), the buffalo (*Bubalus bubalis*) and the bison (*Bison bison*) were aligned with the human myostatin sequence (Accession no. ABI48419.1), and the human myostatin C-terminal domain solved by X-ray diffraction (residues 46-375 out of 375) (**Figure 2A**). *C. dromedarius, C. ferus, C. bactrianus,* and *V. pacos* share 100% of identical amino acids. Six, or fourteen, variants are observed among the above cited *Camelidae* sequences and the human myostatin sequence (Accession No. ABI48419.1), or the corresponding taurine myostatin sequence, respectively (**Figure 2A**). Out of them, 5 in the contrast with the human sequence and 13 in the contrast with the taurine sequence are variants occurring at different sites, while one, at position 164, presented different variants when contrasted with the human and the cattle sequence, respectively. In addition, it is possible to observe that the 6 variants detected in the contrast with the human sequence are conservative (**Figure 2A**), while 9, out of the 14 variants detected in the contrast with the cattle sequence are not conservative (**Figure 2A**). No variants were observed when contrasting among them the sequences from the five species belonging to the *Bovidae* family. A notable exception was *B. bubalis*, for which variants where observed at positions 101, 117 and 141. In all the above cases, the nucleotides observed in *B. bubalis* were different from those observed in all the other eight sequences. **Figure 2B** presents the 3D comparative model of the *C. dromedarius* myostatin dimer, and highlights the variants observed between the *Camelidae* myostatin sequences and the human/bovine myostatin.

Evolutionary Relationships Among Myostatin Proteins

The inferred maximum likelihood tree of myostatin protein sequences (**Figure 3**) highlighted the presence of three supported clusters (bootstrap value higher than 60%), corresponding to Arthropoda, Reptilia and Amphibia, that may reflect a different attitude in the regulation of skeletal muscle growth in the different taxonomic groups. Interestingly, within Mammalian sequences, the highest bootstrap value (99%) was observed for the cluster grouping myostatin sequences belonging to the Bovidae family.

[4]http://genome-euro.ucsc.edu/index.html

[5]https://www.uniprot.org/uniprot/

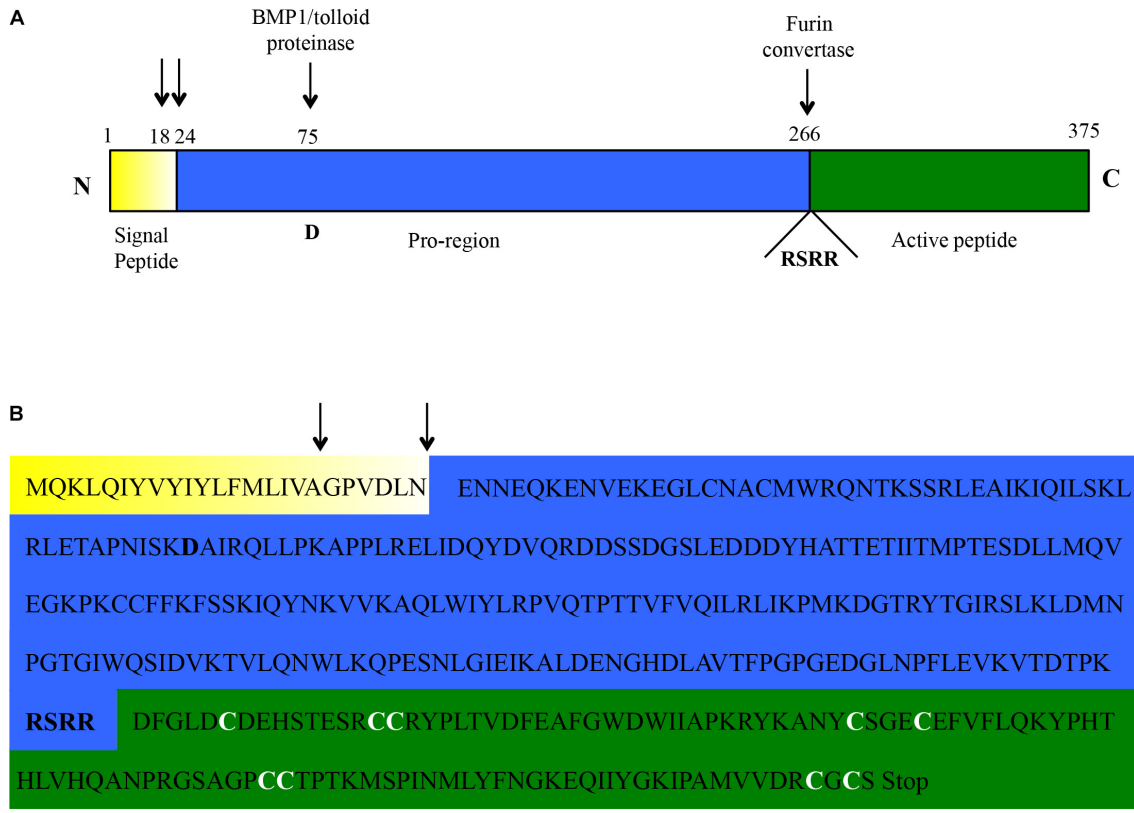


FIGURE 1 | Amino acidic sequence of the *C. dromedarius* myostatin as inferred from the cDNA sequence. **(A)** Schematic outline. The three protein domains (signal peptide, pro-region, and active peptide) are highlighted in different colors (yellow, blue, and green, respectively). The two most likely residues involved in the signal peptide cleavage (see main text) are indicated by black arrows. Similarly, the residue (D, for aspartic acid) shown to be essential for BMP/tolloid protease cleavage, and the motif (RSRR) needed for recognition by furin convertase, are highlighted. **(B)** Amino acidic sequence of the *C. dromedarius* myostatin, with the three protein domains highlighted in different colors, as in **(A)**. The above mentioned hallmarks are also depicted here (signal peptide cleavage, black arrows; BMP/tolloid protease cleavage residue and furin convertase recognition motif, bold). In addition, the nine conserved cysteine residues in the active peptide are indicated (bold and white).

Polymorphism Analysis and *in silico* Functional Prediction

The results of the sequence polymorphism analysis for the myostatin locus are presented in **Table 1**. As can be observed, only three polymorphisms, two transitions (A66460G and T66461C) and one transversion (G66148C), were identified inside the myostatin gene, all located deep in the intron 1. On the other side, a total of 45 and 21 variant sites were identified in our study in the upstream and downstream regions, respectively, with an overall average density of one SNP every about 1.5 kb. In order to interpret the observed patterns of SNP distribution across samples from different countries, we performed a Bayesian clustering analysis of the nine dromedary samples using the 69 identified SNPs. At $K = 7$ (**Supplementary Figure S5**), the analysis highlighted that the sample from Pakistan was well differentiated, and the same occurred for a pair of samples, one from Kingdom of Saudi Arabia and one from United Arab Emirates, respectively. The rest of the samples were clearly "admixed," suggesting that most of the considered SNPs do not follow a geographic pattern. Putative variants in the myostatin coding region, inferred from aligning previously published myostatin sequences with sequences generated in this study, are presented, for completeness, in **Supplementary Figures S4A,B**.

Analysis by Human Splicing Finder highlighted, for the intronic polymorphisms, a potential role in alteration of splicing for G66148C, predicted to break an ESE (exonic splicing enhancer) site (**Supplementary Figure S6A**), A66460G, predicted to generate a new donor site and a new ESS (exonic splicing silencer) site (**Supplementary Figure S6B**), while no significant splicing motif alteration was detected for T66461C (**Supplementary Figure S6C**). TFBS analysis, performed for each SNP using the two input sequences harboring the alternative alleles, highlighted the presence of disrupted TFBSs for all the three loci (**Table 2**). Moreover, we analyzed the potential transcriptional factor binding sites in the DNA sequence of 8 kb of the *C. dromedarius* myostatin gene upstream region (included in the contig-8645394). A total of 10677 putative binding sites were identified (**Supplementary Data Sheet S3**). A graphical outline of the most significant predicted regulatory motifs in the 1.5 kb proximal to the transcription initiation site of the *C. dromedarius* myostatin gene is presented in **Supplementary Figure S7**. In addition, in this region two SNPs were present (T63437C and A64026G) (**Table 1**), out of which

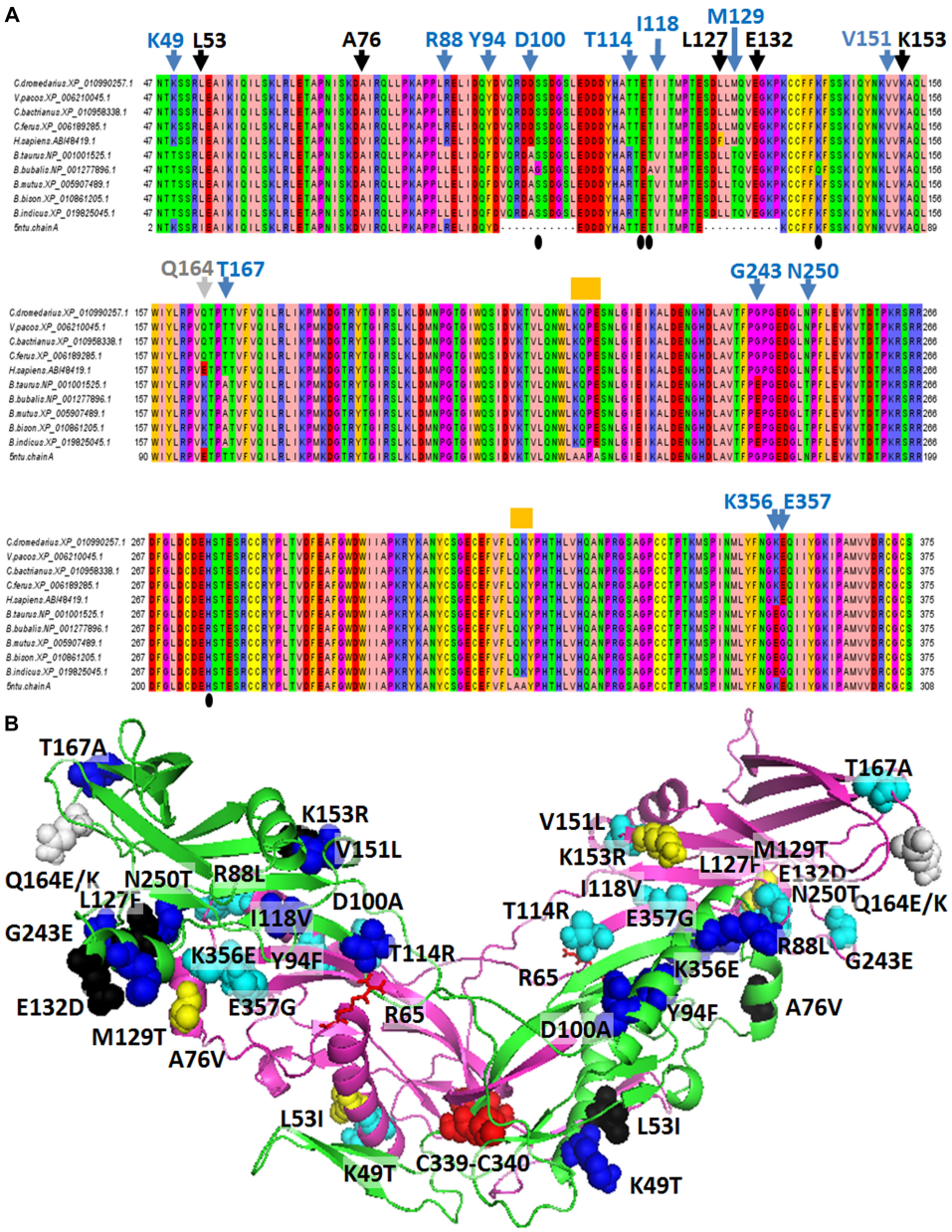


FIGURE 2 | Comparative analysis of myostatin protein sequences. **(A)** The alignment of myostatin orthologous sequences sampled from various mammalian species is presented. Blue arrowheads indicate the variants detected between the four considered *Camelidae* sequences and the five Bovidae sequences at 13 specific sites. Black arrowheads indicate the variants detected between the four considered *Camelidae* sequences and the *Homo sapiens* sequence at 5 specific sites, different from the sites previously cited. The gray arrowhead indicates the position of two different variants detected in *H. sapiens* and in Bovidae in correspondence of Q164 from *C. dromedarius*. Orange "boxes" indicate variations between the *H. sapiens* sequence retrieved from refseq_database and the sequence of the human crystallized myostatin. Amino acid codes and numbering refers to the *C. dromedarius* myostatin. **(B)** Lateral view of the 3D comparative model of the *C. dromedarius* myostatin dimer. The protein is reported in green/magenta cartoon representation. Variants observed between *Camelidae* myostatin sequences and human/Bovidae myostatin are reported in black (5)/blue (13) spheres in chain A, and dark-yellow (5)/cyan (13) spheres in chain B, respectively. The only site of *C. dromedarius* myostatin showing a variation both in *H. sapiens* and in Bovidae locates at site 164 (Q164 for *C. dromedarius*, E164 in *H. sapiens*, K164 in *B. taurus*) and is indicated by gray spheres. Notably, variants observed between *C. dromedarius* and Bovidae occur at different sites with respect to those detected between *C. dromedarius* and *H. sapiens*, with the exclusion of residues at site 164. Residues C339/C340 of chain A and chain B, forming inter-monomer disulphide bridges, are reported in red spheres. R65 of chain A and chain B, involved in interactions with T114, is indicated by red sticks.

the latter was also observed as being polymorphic by aligning the contig-8645394 (**Supplementary Data Sheet S2**) with the publicly available contig4726 (Accession No. JDVD01004726.1) and contig_13989_126 (Accession No. LSZX01094446.1). TFBS analysis, repeated for each SNP using the two input sequences harboring the alternative alleles, suggested the disruption of

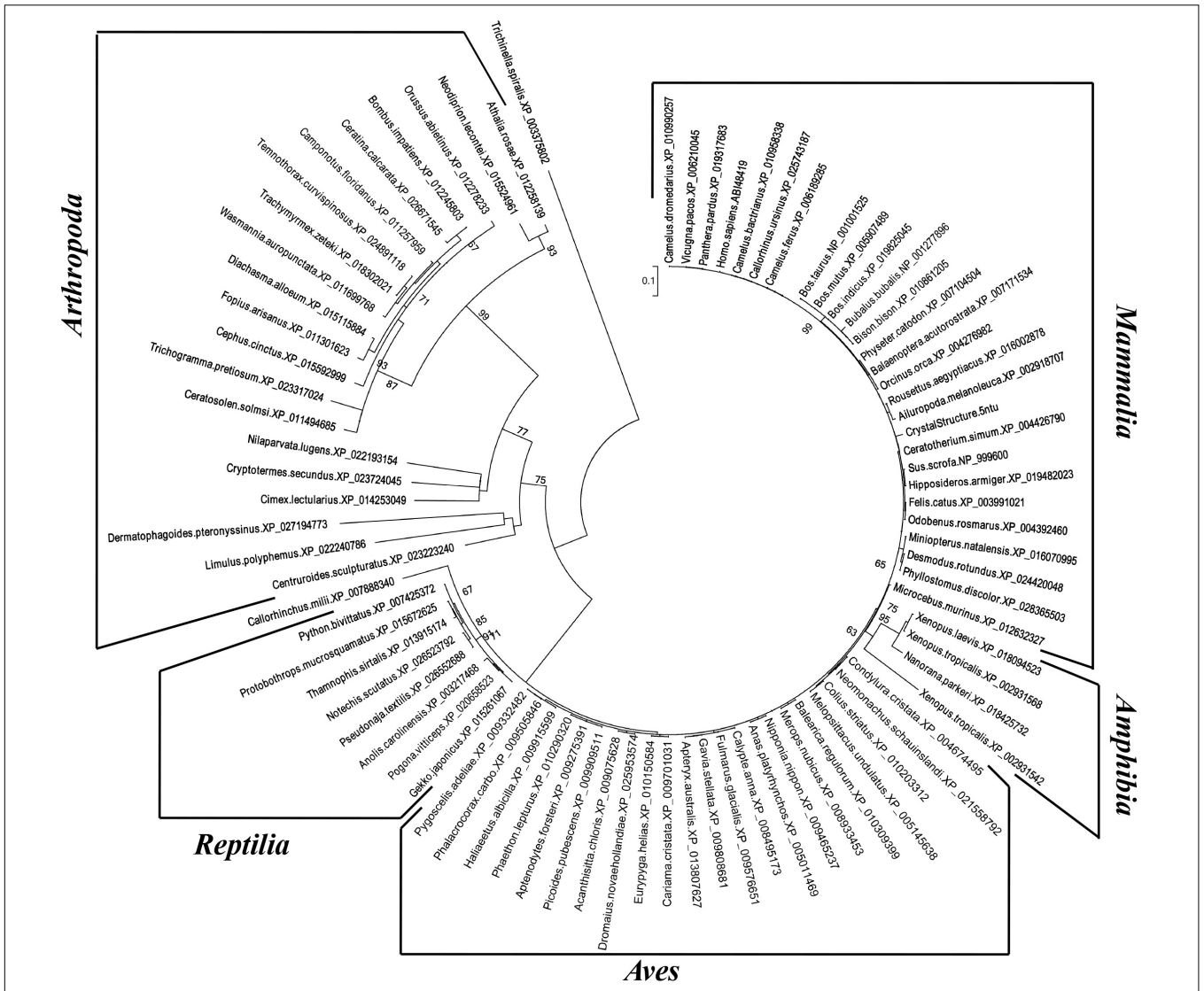


FIGURE 3 | Phylogenetic analysis of myostatin protein sequences. Numbers indicate bootstrap values higher than 60/100.

one (NFKB_Q6) and three (COUP_01, MYB_Q6, and T3R_01) transcription factors binding sites for loci T63437C and A64026G, respectively (data not shown).

Quantitative Myostatin Transcript Analysis in Dromedary Skeletal Muscles

We investigated the quantitative expression of myostatin transcripts in eight dromedary skeletal muscles (*deltoid, extensor carpi radialis, coccygeus, biceps femoris, peroneus longus, semitendinosus, tensor fasciae latae, brachiocephalicus*) by Digital Droplet PCR. Two different assays, with probes targeting the two possible exon junctions in the myostatin gene, were used. Number of droplets generated in each experiment replicate, together with plot of raw data, are shown in **Supplementary Figure S8**. All the replicates passed the cut-off value (>10,000 droplets). **Figure 4** presents the results of the absolute quantification experiments for the eight muscles, and for the two probes, expressed as target copies/μl. The FAM-labeled probe (probe 1, targeting the junction between exon 1 and 2) produced systematically higher values compared to the HEX-labeled probe (probe 2, targeting the junction between exon 2 and 3). However, trends among different muscles were similar for both probes, as also supported by a correlation coefficient higher than 0.78 (data not shown). Analysis of variance among means of different muscles did not find significant difference for any of the two assays.

Myostatin Protein Expression in Dromedary Skeletal Muscles

We investigated the expression of myostatin at the protein level on the same set of dromedary skeletal muscles previously described for quantitative transcript analysis. In order to detect

TABLE 1 | Single Nucleotide Polymorphisms identified in the considered population sample (nine *C. dromedarius* animals from seven countries).

Target region	Size (bp)	Polymorphism
Upstream	64736	G13066T; T13901A°; C14568A°; T14745G; C16352T; G16640A°; G17867A; T25076C°; T29288A; G31182A; G31464A; C32869T°*; C33308G; G33597C; G34295T; T34901C°*; G35625A°*; C35782T°; G38778A; T40104C°*; T41149C*; T42365A; C43439A°; C47473T; C48225T; G48931T°*; C49083T*; G49630C°*; T50414C; T50514G*; A50637C°; T50897C*; T52618C°; G52669T; T53222C°*; G53398A°*; C55949T°*; A56391G; G57208A°*; T58686C°*; G59091T°*; A59187T*; A60587G; T63437C; A64026G°*
5′ UTR	109	None
Exon 1	373	None
Intron 1	1801	G66148C°; A66460G*; T66461C*
Exon 2	374	None
Intron 2	2039	None#
Exon 3	381	None
3′ UTR	215	None
Downstream	21956	A72141T; A72526G; A73136C; G75845A°*; T75862C; G76270A°*; T77354C°*; T78572A; G78774A°; G78993A*; T80531G; C81364T°; C81499T°; G83533C; T86423C°; T87060A°; G89595A°; A89886T; C90627A°; A90686T; T90834C°

Variant sites are presented based on the region where they are located (either in the exons or introns of the myostatin locus, in the 5′ or in the 3′ UTR, or in the upstream/downstream regions). Site positions refer to the C. dromedarius contig-8645394. Circles (°) indicates those SNPs that could also be observed when aligning the contig-8645394 with the publicly available contig4726 (Accession No. JDVD01004726.1). Asterisks (*) indicates those SNPs that could also be observed when aligning the contig-8645394 with the publicly available contig_13989_126 (Accession No. LSZX01094446.1). The # (hash) symbol indicates the only SNP, located in the myostatin gene (intron 2), that was not observed in our study but was evident when aligning the contig-8645394 with both contig4726 and contig_13989_126.

the active C-terminal myostatin dimer, we performed Western Blot analyses using a polyclonal antibody raised against the C-terminus. Moreover, protein electrophoretic separation was run under non-reducing conditions in order to preserve the integrity of the disulphide bonds in the C-terminal domain. As shown in **Figure 5A**, a major band is present at an apparent molecular mass of 75 kDa, corresponding to the expected mass for the promyostatin dimer. Additional, weaker bands were observed at around 40 and 25 kDa, corresponding to the promyostatin monomer and the active C-terminal myostatin dimer, respectively. Densitometric analysis of Western Blots, performed on five different protein extracts for each muscle, highlighted significant differences (ANOVA, $p = 0.0001$) among muscle types for the promyostatin dimer (**Figure 5B**), while no significant difference was observed for both the promyostatin monomer and the active C-terminal myostatin dimer, respectively (**Figure 5B**). *Post hoc* tests highlighted four significant ($p < 0.0017$) pair-wise comparisons, all of them involving the *tensor fasciae latae* muscle (*vs. deltoid*, *extensor carpi radialis, coccygeus*, *and brachiocephalicus*).

DISCUSSION

Myostatin Gene Organization and Protein Comparative Modeling

The myostatin gene has been largely studied in several livestock and model species. Very recently, the gene has been mapped to chromosome 5 in *Camelus dromedarius* (Elbers et al., 2019). By combining experimental and *in silico* approaches, we here describe, for the first time, the gene organization and the protein structure in the one-humped Old World camelid species. We confirmed the major features observed in other species for the myostatin gene. As expected, a high sequence similarity was observed among our experimentally obtained transcript sequence for *C. dromedarius* myostatin and the publicly available predicted sequences for the other two species within the *Camelus* genus, in line with the relatively recent divergence times between them, estimated in 5–8 mya between one-humped and two-humped domestic camels (Wu et al., 2014), and about 0.7 million years ago between *C. bactrianus* and *C. ferus* (Ji et al., 2009). Also, myostatin orthologs showed a high percentage of identical amino acids through *Mammalia*, which may explain the low bootstrap values observed in the maximum likelihood tree. A notable exception was represented by the Bovidae sequences, that clustered with high bootstrap values. Peculiar selection constraints may have played a role in shaping the evolutionary history of myostatin in Bovidae. Indeed, besides the well documented human-mediated positive selection experienced in recent times by the *Bos taurus* myostatin gene, particularly in specialized beef breeds, evidence for a more remote action of positive selection on this gene, operating during the time of divergence of Bovinae and Antilopinae, has been produced by Tellgren et al. (2004) in a systematic analysis of myostatin sequence evolution in ruminants. These periods of positive selective pressure on myostatin may correlate with changes in skeletal muscle mass. In fact, the early bovid fossil record, dating back to around 17 million years ago, had a body mass estimate of only around 20 kg. Hence, the hypothesis is that selective pressures on myostatin drove this increase in body mass coupled to an increase in skeletal muscle mass, in turn driven by ecological changes in the environments of the various species. In a recent paper, the phylogenetic relationships between camelids and other mammalian species were investigated using whole-genome sequence data (Wu et al., 2014). The authors report estimated

divergence time between camelids and cattle lower than those between other mammals. This seeming discrepancy with our results may arise from the fact that single gene phylogenies could not reflect the complex evolutionary history of a whole genome and they could be less reliable in inferring genome-scale events.

Myostatin protein consists of a non-covalently held complex of the N-terminal propeptide and a disulfide-linked dimer of C-terminal fragments (Lee and McPherron, 2001). Variants detected respectively in *B. taurus* and *H. sapiens*, in correspondence of the *C. dromedarius* K49 and L53 may be involved in the stability of the myostatin dimer due to their location close to the disulphide bonds occurring among C339 and C340 residues at the myostatin dimer interface (Jiang et al., 2004). Notably, the myostatin N-terminal domain contains a region (residues 49–67) highly similar to the key latency-determining regions of the TGF-β superfamily (Walton et al., 2010). Thus, it is expected that variations observed in our comparative analysis at this region, above all in correspondence of the *C. dromedarius* K49 (Thr in *B. taurus* myostatin, Lys in TGF-β1) and L53 (Ile in *H. sapiens* myostatin, Leu in TGF-β1) may contribute toward its latency, according to Walton et al. (2010). More in general, all the cited variants locate at the monomer/monomer interface. Notably, we observed, in *Camelidae* myostatin sequences, not conservative substitutions compared to the *B. taurus taurus* myostatin sequence at position 49 (Lys in *C. dromedarius*; Thr in *B. taurus*), 88 (Arg in *C. dromedarius*; Leu in *B. taurus*), 100 (Asp in *C. dromedarius*; Ala in *B. taurus*),114 (Thr in *C. dromedarius*; Arg in *B. taurus*), 129 (Met in *C. dromedarius*, Thr in *B. taurus*), 167 (Thr in *C. dromedarius*, Ala in *B. taurus*), 243 (Gly in *C. dromedarius*; Glu in *B. taurus*), 356 (Lys in *C. dromedarius*; Glu in *B. taurus*), 357 (Glu in *C. dromedarius*, Gly in *B. taurus*) that may produce a different charge network in myostatin dimer, favoring different monomer/monomer interactions. In particular, among the described variants, it is worth noting that the residue at site 114 forms intra-chain binding interactions with Y111 and H112 and inter-chain binding interactions with R65. R65, Y111 and H112 together with K153 (R153 in *H. sapiens*) were already described as "fastener" residues associated with muscle- and obesity-related

TABLE 2 | Results of the TFBS analysis for the three intronic SNPs detected in this study.

SNP	AC	ID	Score	Strand	Consensus	Signal
A66460G					**Allele A**	
	M00103	V$CLOX_01	0.778624	(+)	NNTATCGATTANYNW	GGTATTAATTAGCTG
	M00104	V$CDPCR1_01	0.790948	(–)	NATCGATCGS	GGTATTAATT
	M00134	V$HNF4_01	0.769283	(–)	NNNRGGNCAAAGKTCANNN	ATTTAAATTTTGGTATTAA
					Allele G	
	M00211	V$PADS_C	0.825857	(+)	NGTGGTCTC	TTTGGTGTT
	M00212	V$POLY_C	0.787056	(+)	CAATAAAACCYYYYKCTN	CATTTAAATTTTGGTGTT
	M00279	V$MIF1_01	0.741931	(–)	NNGTTGCWWGGYAACNGS	GGTGTTAATTAGCTGCTA
	M00280	V$RFX1_01	0.776836	(–)	NNGTNRCNWRGYAACNN	GTGTTAATTAGCTGCTA
G66148C					**Allele G**	
	M00143	V$PAX5_01	0.787091	(+)	NCNNNRNKCANNGNWGNRKRGCSRSNNN	GAGACAGGCACCTTAACAGAGAAGGCAT
					Allele C	
	M00262	V$STAF_01	0.767919	(+)	NTTWCCCANMATGCAYYRCGNY	TTAACACAGAAGGCATGACAAG
T66461C					**Allele T**	
	M00103	V$CLOX_01	0.778624	(+)	NNTATCGATTANYNW	GGTATTAATTAGCTG
	M00104	V$CDPCR1_01	0.790948	(–)	NATCGATCGS	GGTATTAATT
	M00252	V$TATA_01	0.829231	(+)	STATAAAWRNNNNNN	GTATTAATTAGCTGC
					Allele C	
	M00185	V$NFY_Q6	0.779548	(–)	TRRCCAATSRN	CTAATTAGCTG

AC, transcription factor matrix code. ID, transcription factor label, with V meaning "vertebrate." SCORE, similarity (0.0–1.0) between a registered sequence for the transcription factor binding sites and the input sequence. STRAND, strandness. + and – means forward and reverse strands that the transcription factor binds, respectively. CONSENSUS, consensus sequence (fixed) of the transcription factor binding sites. S = C or G, W = A or T, R = A or G, Y = C or T, K = G or T, M = A or C, N = any base pair. SIGNAL, sub-sequence from the input sequence at the position corresponding to the consensus sequence. Default cut-offs by Tsunoda and Takagi, 1999 were adopted. The analysis was performed, for each SNP, using the two input sequences harboring the alternative alleles.

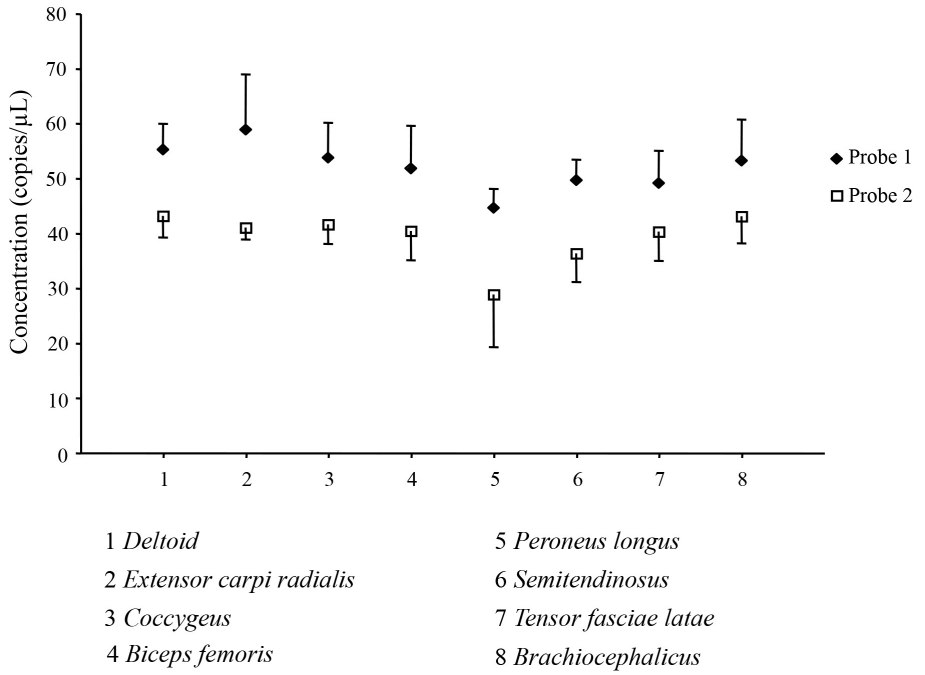


FIGURE 4 | Absolute quantification of the myostatin transcript in skeletal muscles. Results of the Digital Droplet PCR analysis for the eight considered dromedary muscles (1, *deltoid*; 2, *extensor carpi radialis*; 3, *coccygeus*; 4, *biceps femoris*; 5, *peroneus longus*; 6, *semitendinosus*; 7, *tensor fasciae latae*; 8, *braciocephalicus*) and for the two used probes (u, Probe 1; ◆, □ Probe 2) are presented as mean ± SD of the four replicates.

phenotypes (Gonzalez-Freire et al., 2010; Santiago et al., 2011; Bhatt et al., 2012; Garatachea et al., 2013; Szlama et al., 2015). Furthermore, variations of the residue at site 100 may influence the release of the active form from the myostatin propeptide complex due to its location close to the myostatin propeptide TLD cleavage target site (consisting of the dipeptide 98-RD-99) (Szlama et al., 2013).

Genetic Sequence Polymorphism and Functional Prediction

Unexpectedly, the Next Generation Sequencing (NGS) of the whole myostatin locus in nine dromedaries from a variety of geographic locations in Asia, Africa and Europe did not allow the identification of novel intra-genic variants and only confirmed the presence of the three SNPs in intron 1 previously identified by Muzzachi et al. (2015) via Sanger-sequencing of a reduced portion of the gene on a different set of Northern African dromedaries. This result seems to support the hypothesis, formulated by Muzzachi et al. (2015) that the low diversity observed at the myostatin locus in *Camelus dromedarius* may reflect the peculiar evolutionary history of this species, which likely developed as domesticates from a low variable wild ancestor population.

Evidence about the existence of functional variants located in introns is growingly accumulating (Lee et al., 2015; Mou et al., 2015; Hong et al., 2018; Ostrovsky et al., 2018), not only restricted to exon–intron boundaries but also in deep intronic regions (Mendes de Almeida et al., 2017; Vaz-Drago et al., 2017). The three intronic SNPs detected in this study were *in silico* predicted to have the potential of altering both splicing and TFBS, thus suggesting they may play a role in myostatin processing and/or regulation. TFBS analysis in the myostatin upstream region allowed to predict several potential transcriptional factor binding sites, mainly belonging to the large family of dimerizing transcription factors harboring a basic helix-loop-helix (bHLH) structural motif, such as MYOD (myogenic differentiation), MYOG (myogenin), MYC (myelocytomatosis viral oncogene), MAX (MYC Associated Factor X), TAL1 (T-Cell Acute Lymphocytic Leukemia), SREBP (Sterol Regulatory Element-Binding Protein), AHR (Aryl Hydrocarbon Receptor), ARNT (Aryl hydrocarbon receptor nuclear translocator), HEN (Nescient Helix-Loop-Helix 1), HLF (Hepatic Leukemia Factor), USF (Upstream Transcription Factor) (Jones, 2004; Sailsbery and Dean, 2012). Out of them, MYOD and MYOG belong to the myogenic regulatory factor (MRF) family known to play key roles in the determination and differentiation of skeletal muscle (Botzenhart et al., 2018). Besides them, a role in regulating myostatin expression has been largely demonstrated for CREB (Xie et al., 2018), MEF (Bo Li et al., 2012; Estrella et al., 2015) and C/EBP (Allen et al., 2010; Deng et al., 2012), for which potential binding sites were also detected in our *in silico* analysis of the dromedary myostatin gene upstream sequence. Moreover, in the about 400 bp region upstream to the transcriptional start site, three TATA boxes and one CCAAT box were observed, consistently with previous reports

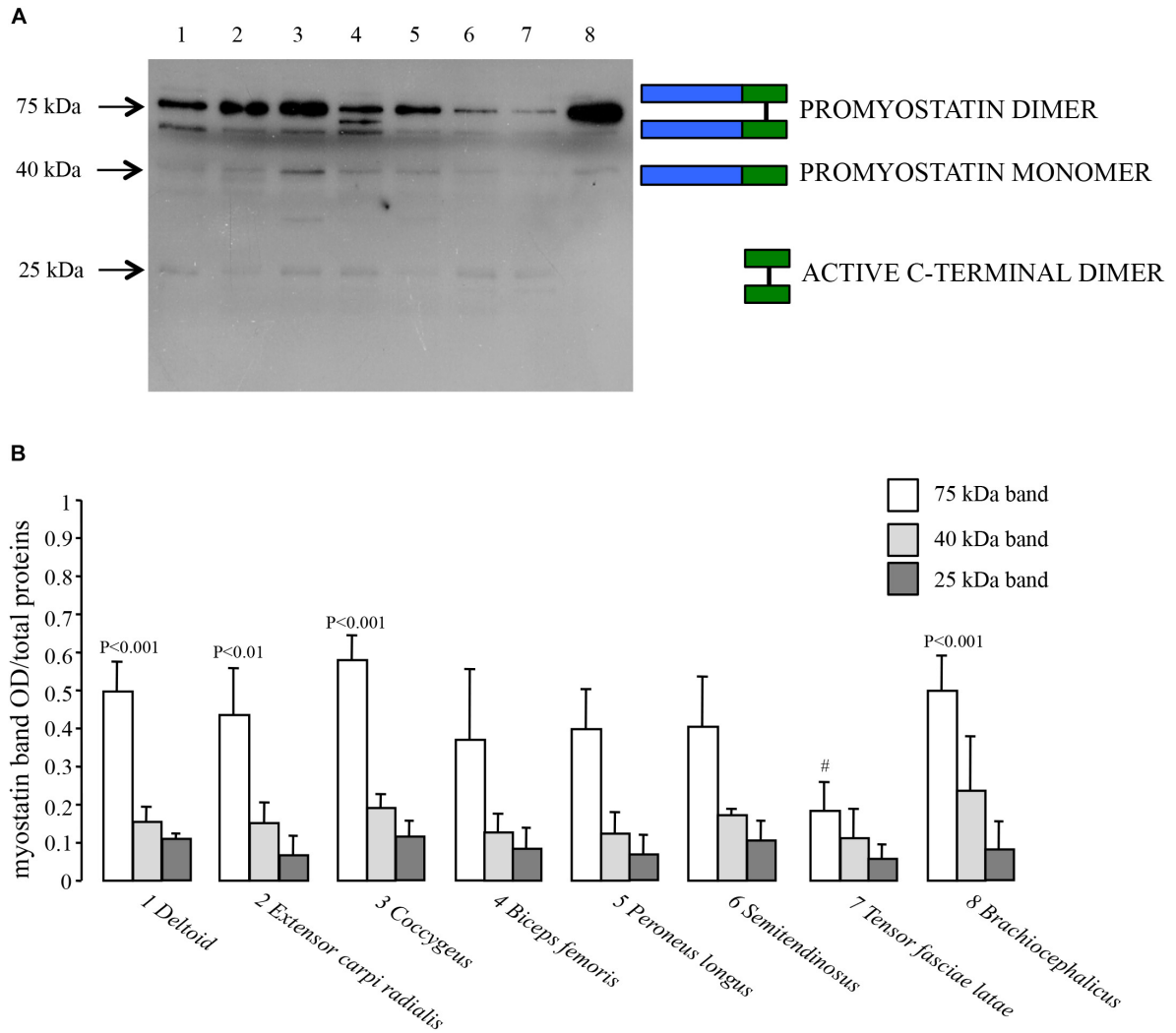


FIGURE 5 | Myostatin protein expression. **(A)** Representative Western Blot of the eight considered dromedary muscles (1, *deltoid*; 2, *extensor carpi radialis*; 3, *coccygeus*; 4, *biceps femoris*; 5, *peroneus longus*; 6, *semitendinosus*; 7, *tensor fasciae latae*; 8, *brachiocephalicus*) performed using a rabbit polyclonal antibody (1:1000 dilution) that specifically binds the carboxy-terminal domain (Origene, TA343358). **(B)** Western Blot densitometric analysis of the promyostatin dimer (75 kDa), the promyostatin monomer (40 kDa) and the active C-terminal dimer (25 kDa), respectively. For each lane, optical density (OD) of the considered band is presented as a ratio over the total density of all the proteins transferred on the membrane and stained with Coomassie blue.

by Spiller et al. (2002) in the bovine species and Du et al. (2011) in the ovine species. On the contrary, in the above region, we detected four E-boxes, unlike (Spiller et al., 2002) and (Du et al., 2005), who detected three and five E-boxes, respectively.

Most of the variants detected in our population sample, representative of seven different countries across the African and the Euro/Asiatic continent, could be also detected by aligning our *Camelus dromedarius* contig-8645394 with contigs available in public databases, representative of animals of African and Asiatic descent (Accession No. LSZX01094446.1, from a Targui animal sampled in Algeria, and Accession No. JDVD01004726, from an animal sampled in the Kingdom of Saudi Arabia, respectively). This result suggests that (i) our population sample, despite being limited in size, could be considered as providing a good representation of the species genetic diversity, and (ii) a limited differentiation may exist also among geographically distant samples, as pointed out by the preliminary results of the first world-wide *Camelus dromedarius* genetic diversity survey performed using genome-wide RAD-sequencing (Ciani et al., 2017)[6], and in line with the known recurrent gene flow at ancient trading centers along the caravan routes (Almathen et al., 2016).

Myostatin Expression Profiling in Skeletal Muscles

In order to quantify the level of myostatin transcript in dromedary skeletal muscle, we adopted a Digital Droplet

[6]https://pag.confex.com/pag/xxv/meetingapp.cgi/Paper/23709

PCR approach. In our experiments, we used two different probes designed on the two exon-junctions of the myostatin gene and differentially dye-labeled. The FAM-labeled probe gave systematically higher values compared to the HEX-labeled one. This result agrees with the known evidence that FAM has a stronger signal compared to other dyes. Indeed, this feature may determine a larger number of droplets, where target amplification occurred for both assays, to be designed as "positive" for the FAM-labeled probe compared to the HEX-labeled probe at a given fluorescence threshold. Alternatively, a different absolute quantification in a two-probe system targeting the same gene may arise from the presence of alternative transcripts that may reduce the amplification/detection efficiency of one of the two assays, but not necessarily both. However, in our study, given the systematic and consistent differences between the two assays across eight different skeletal muscles, a major role for the alternative transcript phenomenon appears rather unlike. Finally, we cannot exclude a minor role of stochastic factors in affecting the observed results.

In our experimental conditions, no significant differences were observed for the eight considered muscles by any of the two assays. These results are in line with those previously published by Morrison et al. (2014) who did not find significant variation in myostatin expression levels, assessed via Quantitative Real-time PCR, among four skeletal muscles (*rectus abdominis, longus colli, adductor, pectoralis transversus*) in the horse species. A similar scenario was observed in this study also at the protein level where no significant difference was observed among muscle types for the three myostatin forms (25, 40, and 75 kDa), with the exception of four pair-wise comparisons, all involving the *tensor fasciae latae* muscle, where a significantly lower expression was observed only for the promyostatin dimer. For the other two myostatin forms, *tensor fasciae latae* displayed weaker bands although they did not reach significance when contrasted to other muscles. The generally low expression of the three myostatin forms in the *tensor fasciae latae* muscle tempted us to speculate about a possible relationship with the fiber type composition of this muscle, which has been reported, in various species, to be mainly of the fast glycolytic type (Ariano et al., 1973; Abe et al., 1987; Manabe et al., 2000; Sazili et al., 2005; Bakou et al., 2015). In fact, some authors previously reported about a negative correlation between myostatin and the fast phenotype of skeletal muscles (Bouley et al., 2005; Hennebry et al., 2009; Baan et al., 2013). However, muscle phenotypes may be affected by many endogenous and exogenous factors, such as stage of maturity (Kugelberg, 1976), level of activity (Goldspink, 1983), different sampling regions of the same muscle (Torrella et al., 2000), histological method (Karlsson et al., 1999). Hence, the known large variability of muscle fiber phenotypes, coupled to the lack of specific data for the dromedary camels, makes it hazardous to extend the mentioned correlation to the species under study.

In general, densitometric analysis highlighted that the promyostatin dimer is the most expressed form in all the considered muscles while the active myostatin has the lowest level of expression. The above results fit well with multiple evidences that, in muscle, myostatin resides primarily as unprocessed promyostatin (Anderson et al., 2008; Pirruccello-Straub et al., 2018) and that the active mature growth factor is significantly less abundant in this compartment (Hill et al., 2002, 2003; Zimmers et al., 2002; Anderson et al., 2008; Lakshman et al., 2009). Moreover, it must be pointed out that the observed 25 kDa bands, suggestive of the active myostatin form, could, in our study, reflect the amount of myostatin dimers, deriving from the latent complex generated by furin cleavage, and artificially "activated" by the experimental SDS environment (Wehling et al., 2000), rather than reflecting a physiologically activated form. Based on the above, the unprocessed or partially processed myostatin dimers could act as important reservoirs of slowly available myostatin forms, and the sequential cleavage steps contribute an additional layer of control, within an already complex regulatory framework.

ETHICS STATEMENT

Tissue sampling was done on slaughterhouses from dead animals. Blood sampling from 25 Old World camels was collected during routine veterinary procedures or as part of a monitoring program of the wild camel population in Mongolia.

AUTHOR CONTRIBUTIONS

MF performed the experiments, analyzed the data, prepared the figures and contributed toward writing the manuscript. RF performed the experiments and analyzed the data. LG and CP performed the experiments, analyzed the data, and contributed to the discussion and toward writing the manuscript. BF helped in sample collection and contributed to the discussion. AO performed the RACE-PCR experiments and contributed to the discussion. PB performed the experiments, analyzed the data, and contributed to the discussion. EC designed the research, performed the experiments, analyzed the data, and wrote the manuscript. All authors have read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00502/full#supplementary-material

FIGURE S1 | Electropherograms of 5' and 3' RACE PCR products. **(A)** Partial sequence of the 5' untranslated region (UTR) of the *C. dromedarius* myostatin transcript, together with a partial sequence of the universal oligonucleotide provided in the SMARTer RACE cDNA amplification kit (Clonetech). **(B)** Partial sequence of the 3' untranslated region (UTR) of the *C. dromedarius* myostatin transcript, together with part of the poly(A) tail.

FIGURE S2 | RACE PCR products separated via 2% agarose gel electrophoresis. The upper arrow corresponds to 366 base pairs, while the lower arrow corresponds to 316 base pairs.

FIGURE S3 | Comparative analysis of the myostatin genomic locus in the three Old World camelid species. Dromedary, *C. dromedarius* (contig-8645394). Wild camel, *C. ferus* (contig-7907533). Bactrian camel, *C. bactrianus* (contig-8938518). Highlighted, in the *C. dromedarius* sequence, the 5' and 3' UTR regions (light blue) and the three exons (yellow).

FIGURE S4 | Putative SNPs in the myostatin coding region, as inferred from alignment among previously published myostatin sequences and sequences generated in this study. **(A)** Consensus sequence of the myostatin coding region. Highlighted, in yellow, the exonic region, and, in red, the variant sites. Numbering of polymorphism positions refers to the contig AGVR01040332. **(B)** Summary table of the inferred polymorphisms showing the nature of the putative variants, together with the reference literature, and the prediction of the variant effects (in case of a missense mutation, the alternative amino acids are reported, otherwise "none" is entered).

FIGURE S5 | Plot of the Bayesian clustering analysis perfoemed on the nine considered dromedaries using the 69 identified SNPs. The plot shows the results obtained for K = 7, that was identified as the most likely output by visual inspection of the probability values associated to each tested K value (from 1 to 9). Numbers indicate different samples (1, United Arab Emirates; 2, Qatar; 3-4-5, Kingdom of Saudi Arabia; 6, Austria; 7, Kenya; 8, Sudan; 9, Pakistan). Colors indicate the seven different clusters. The proportion of each individual sample in each inferred cluster is shown in the y-axis.

FIGURE S6 | Results of the Human Splicing Finder analysis for the three intronic SNPs detected in this study. **(A)** G66148C. **(B)** A66460G. **(C)** T66461C.

FIGURE S7 | Graphical outline of the major predicted regulatory motifs in the 1.5 kb proximal to the transcription initiation site of the *C. dromedarius* myostatin gene. TATA boxes (highlighted in yellow), E-boxes (highlighted in gray), CREBP1 (light green), CREB_01 (orange), MYOGNF1 (blue), CEBP_01 (purple), MEF (pink), MYOD (dark green), and the CCAAT box (highlighted in purple) are presented.

FIGURE S8 | Graphical evaluation of the Digital Droplet PCR performance. **(A)** 2D-dot plot of fluorescent signals detected in the Digital Droplet PCR experiments. Fluorescence results are plotted as two-dimensional dot plots. Gray dots correspond to empty droplets. Blue dots correspond to droplets containing at least one copy of the sequence complementary to Probe 1. Green dots correspond to droplets containing at least one copy of the sequence complementary to Probe 2. Orange dots correspond to droplets containing at least one copy of the sequence complementary to Probe 1 and at least one copy of the sequence complementary to Probe 2 (double-positive droplets). The observed pattern is relatively well balanced and dot clouds are well separated, suggesting (i) the absence of significant bias in the amplification of the two regions targeted by the two considered probes (exon1/exon2, and exon2/exon3) and (ii) the applicability of the BIORAD proprietary auto-analysis tool for target quantification. **(B)** Number of droplets generated in the Digital Droplet PCR experiments. Results for eight dromedary skeletal muscles, each from two different animals (biological replicate), analyzed in duplicate (technical replicate) are shown.

TABLE S1 | Sequence accession numbers (left panel) and links to the webpages of databases and softwares (right panel) considered in this study.

DATA SHEET S1 | Amino acidic sequences of the 83 non redundant myostatin proteins used for building the maximum likelihood tree presented in **Figure 3**. The data are in the "multiple sequence alignment" format.

DATA SHEET S2 | Sequence of the *Camelus dromedarius* contig-8645394. Exons (highlighted in yellow), 5' and 3' UTR regions (underlined), SNPs (highlighted in red), the GAT codon for the aspartic acid essential for BMP/tolloid protease cleavage (highlighted in green), and the 12 nucleotides coding for the RSRR motif (highlighted in gray), needed for recognition by furin convertase, are shown.

DATA SHEET S3 | Results of the TFBIND analysis for the SNPs present in the *C. dromedarius* myostatin gene upstream region (8 kb). AC, transcription factor matrix code. ID, transcription factor label, with V meaning "vertebrate". SCORE, similarity (0.0–1.0) between a registered sequence for the transcription factor binding sites and the input sequence. STRAND, strandness. + and − means forward and reverse strands that the transcription factor binds, respectively. CONSENSUS, consensus sequence (fixed) of the transcription factor binding sites. S = C or G, W = A or T, R = A or G, Y = C or T, K = G or T, M = A or C, N = any base pair. SIGNAL, sub-sequence from the input sequence at the position corresponding to the consensus sequence. Default cut-offs by Tsunoda and Takagi, 1999 were adopted. The analysis was performed, for each SNP, using the two input sequences harboring the alternative alleles.

REFERENCES

Abe, J., Fujii, Y., Kuwamura, Y., and Hizawa, K. (1987). Fiber type differentiation and myosin expression in regenerating rat muscles. *Acta Pathol. Jpn.* 37, 1537–1547. doi: 10.1111/j.1440-1827.1987.tb02465.x

Agrawal, V., Gahlot, G. C., Ashraf, M., Khicher, J. P., and Thakur, S. (2017). Sequence analysis and phylogenetic relationship of myostatin gene of bikaneri Camel (*Camelus dromedarius*). *J. Camel Pract. Res.* 24:73. doi: 10.5958/2277-8934.2017.00011.X

Aiello, D., Patel, K., and Lasagna, E. (2018). The myostatin gene: an overview of mechanisms of action and its relevance to livestock animals. *Anim. Genet.* 49, 505–519. doi: 10.1111/age.12696

Allen, D. L., Bandstra, E. R., Harrison, B. C., Thorng, S., Stodieck, L. S., Kostenuik, P. J., et al. (2009). Effects of spaceflight on murine skeletal muscle gene expression. *J. Appl. Physiol.* 106, 582–595. doi: 10.1152/japplphysiol.90780.2008

Allen, D. L., Cleary, A. S., Hanson, A. M., Lindsay, S. F., and Reed, J. M. (2010). CCAAT/enhancer binding protein-delta expression is increased in fast skeletal muscle by food deprivation and regulates myostatin transcription in vitro. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 299, R1592–R1601. doi: 10.1152/ajpregu.00247.2010

Allen, D. L., Cleary, A. S., Speaker, K. J., Lindsay, S. F., Uyenishi, J., Reed, J. M., et al. (2008). Myostatin, activin receptor IIb, and follistatin-like-3 gene expression are altered in adipose tissue and skeletal muscle of obese mice. *Am. J. Physiol. Endocrinol. Metab.* 294, E918–E927. doi: 10.1152/ajpendo.00798.2007

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-terWengel, P., Pitt, D., et al. (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6707–6712. doi: 10.1073/pnas.1519508113

Anderson, S. B., Goldberg, A. L., and Whitman, M. (2008). Identification of a novel pool of extracellular pro-myostatin in skeletal muscle. *J. Biol. Chem.* 283, 7027–7035. doi: 10.1074/jbc.M706678200

Ariano, M. A., Armstrong, R. B., and Edgerton, V. R. (1973). Hindlimb muscle fiber populations of five mammals. *J. Histochem. Cytochem.* 21, 51–55. doi: 10.1177/21.1.51

Baan, J. A., Kocsis, T., Keller-Pinter, A., Muller, G., Zador, E., Dux, L., et al. (2013). The compact mutation of myostatin causes a glycolytic shift in the phenotype of fast skeletal muscles. *J. Histochem. Cytochem.* 61, 889–900. doi: 10.1369/0022155413503661

Bakou, S. N., Nteme Ella, G. S., Aoussi, S., Guiguand, L., Cherel, Y., and Fantodji, A. (2015). Fiber composition of the grasscutter (*Thryonomys swinderianus*. Temminck 1827) thigh muscle: an enzyme-histochemical study. *J. Cytol. Histol.* 6:311. doi: 10.4172/2157-7099.1000311

Bhatt, S. P., Nigam, P., Misra, A., Guleria, R., Luthra, K., Jain, S. K., et al. (2012). Association of the Myostatin gene with obesity, abdominal obesity and low lean body mass and in non-diabetic Asian Indians in north India. *PLoS One* 7:e40977. doi: 10.1371/journal.pone.0040977

Bo Li, Z., Zhang, J., and Wagner, K. R. (2012). Inhibition of myostatin reverses muscle fibrosis through apoptosis. *J. Cell Sci.* 125(Pt 17), 3957–3965. doi: 10.1242/jcs.090365

Botzenhart, U. U., Gerlach, R., Gredes, T., Rentzsch, I., Gedrange, T., and Kunert-Keil, C. (2018). Expression rate of myogenic regulatory factors and muscle growth factor after botulinum toxin A injection in the right masseter muscle of dystrophin deficient (mdx) mice. *Adv. Clin. Exp. Med.* 28, 11–18. doi: 10.17219/acem/76263

Bouley, J., Meunier, B., Chambon, C., De Smet, S., Hocquette, J. F., and Picard, B. (2005). Proteomic analysis of bovine skeletal muscle hypertrophy. *Proteomics* 5, 490–500. doi: 10.1002/pmic.200400925

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254. doi: 10.1006/abio.1976.9999

Carlson, C. J., Booth, F. W., and Gordon, S. E. (1999). Skeletal muscle myostatin mRNA expression is fiber-type specific and increases during hindlimb unloading. *Am. J. Physiol.* 277(Pt 2), R601–R606.

Cash, J. N., Rejon, C. A., McPherron, A. C., Bernard, D. J., and Thompson, T. B. (2009). The structure of myostatin:follistatin 288: insights into receptor utilization and heparin binding. *EMBO J.* 28, 2662–2676. doi: 10.1038/emboj.2009.205

Cornelison, D. D., Olwin, B. B., Rudnicki, M. A., and Wold, B. J. (2000). MyoD(-/-) satellite cells in single-fiber culture are differentiation defective and MRF4 deficient. *Dev. Biol.* 224, 122–137. doi: 10.1006/dbio.2000.9682

Cotton, T. R., Fischer, G., Wang, X., McCoy, J. C., Czepnik, M., Thompson, T. B., et al. (2018). Structure of the human myostatin precursor and determinants of growth factor latency. *EMBO J.* 37, 367–383. doi: 10.15252/embj.201797883

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Deng, B., Wen, J., Ding, Y., Gao, Q., Huang, H., Ran, Z., et al. (2012). Functional analysis of pig myostatin gene promoter with some adipogenesis- and myogenesis-related factors. *Mol. Cell. Biochem.* 363, 291–299. doi: 10.1007/s11010-011-1181-y

Desmet, F. O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., and Beroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37:e67. doi: 10.1093/nar/gkp215

Du, R., Chen, Y. F., An, X. R., Yang, X. Y., Ma, Y., Zhang, L., et al. (2005). Cloning and sequence analysis of myostatin promoter in sheep. *DNA Seq.* 16, 412–417. doi: 10.1080/10425170500226474

Du, R., Du, J., Qin, J., Cui, L.-C., Hou, J., Guan, H. et al. (2011). Molecular cloning and sequence analysis of the cat myostatin gene 5′ regulatory region. *Afr. J. Biotechnol.* 10, 10366–10372. doi: 10.5897/AJB10.2577

Dunner, S., Miranda, M. E., Amigues, Y., Canon, J., Georges, M., Hanset, R., et al. (2003). Haplotype diversity of the myostatin gene among beef cattle breeds. *Genet. Sel. Evol.* 35, 103–118. doi: 10.1051/gse:2002038

Elbers, J. P., Rogers, M. F., Perelman, P. L., Proskuryakova, A. A., Serdyukova, N. A., Johnson, W. E., et al. (2019). Improving illumina assemblies with Hi-C and long reads: an example with the North African dromedary. *Mol. Ecol. Resour.* doi: 10.1111/1755-0998.13020 [Epub ahead of print].

Elkina, Y., von Haehling, S., Anker, S. D., and Springer, J. (2011). The role of myostatin in muscle wasting: an overview. *J. Cachexia Sarcopenia Muscle* 2, 143–151. doi: 10.1007/s13539-011-0035-35

Estrella, N. L., Desjardins, C. A., Nocco, S. E., Clark, A. L., Maksimenko, Y., and Naya, F. J. (2015). MEF2 transcription factors regulate distinct gene programs in mammalian skeletal muscle differentiation. *J. Biol. Chem.* 290, 1256–1268. doi: 10.1074/jbc.M114.589838

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Ferrell, R. E., Conte, V., Lawrence, E. C., Roth, S. M., Hagberg, J. M., and Hurley, B. F. (1999). Frequent sequence variation in the human myostatin (GDF8) gene as a marker for analysis of muscle-related phenotypes. *Genomics* 62, 203–207. doi: 10.1006/geno.1999.5984

Gao, F., Kishida, T., Ejima, A., Gojo, S., and Mazda, O. (2013). Myostatin acts as an autocrine/paracrine negative regulator in myoblast differentiation from human induced pluripotent stem cells. *Biochem. Biophys. Res. Commun.* 431, 309–314. doi: 10.1016/j.bbrc.2012.12.105

Garatachea, N., Pinos, T., Camara, Y., Rodriguez-Romo, G., Emanuele, E., Ricevuti, G., et al. (2013). Association of the K153R polymorphism in the myostatin gene and extreme longevity. *Age* 35, 2445–2454. doi: 10.1007/s11357-013-9513-9513

Goldspink, G. (1983). "Alterations in myofibril size and structure during growth, exercise, and changes in environmental temperature," in eds L. D. Peachey, R. H. Adrian, and S. R. Geiger (Baltimore: Williams Wilkins), 539–554.

Gonzalez-Cadavid, N. F., Taylor, W. E., Yarasheski, K., Sinha-Hikim, I., Ma, K., Ezzat, S., et al. (1998). Organization of the human myostatin gene and expression in healthy men and HIV-infected men with muscle wasting. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14938–14943. doi: 10.1073/pnas.95.25.14938

Gonzalez-Freire, M., Rodriguez-Romo, G., Santiago, C., Bustamante-Ara, N., Yvert, T., Gomez-Gallego, F., et al. (2010). The K153R variant in the myostatin gene and sarcopenia at the end of the human lifespan. *Age* 32, 405–409. doi: 10.1007/s11357-010-9139-9137

Gustafsson, T., Osterlund, T., Flanagan, J. N., von Walden, F., Trappe, T. A., Linnehan, R. M., et al. (2010). Effects of 3 days unloading on molecular regulators of muscle size in humans. *J. Appl. Physiol.* 109, 721–727. doi: 10.1152/japplphysiol.00110.2009

Guttridge, D. C., Mayo, M. W., Madrid, L. V., Wang, C. Y., and Baldwin, A. S. Jr. (2000). NF-kappaB-induced loss of MyoD messenger RNA: possible role in muscle decay and cachexia. *Science* 289, 2363–2366. doi: 10.1126/science.289.5488.2363

Hennebry, A., Berry, C., Siriett, V., O'Callaghan, P., Chau, L., Watson, T., et al. (2009). Myostatin regulates fiber-type composition of skeletal muscle by regulating MEF2 and MyoD gene expression. *Am. J. Physiol. Cell Physiol.* 296, C525–C534. doi: 10.1152/ajpcell.00259.2007

Hill, J. J., Davies, M. V., Pearson, A. A., Wang, J. H., Hewick, R. M., Wolfman, N. M., et al. (2002). The myostatin propeptide and the follistatin-related gene are inhibitory binding proteins of myostatin in normal serum. *J. Biol. Chem.* 277, 40735–40741. doi: 10.1074/jbc.M206379200

Hill, J. J., Qiu, Y., Hewick, R. M., and Wolfman, N. M. (2003). Regulation of myostatin in vivo by growth and differentiation factor-associated serum protein-1: a novel protein with protease inhibitor and follistatin domains. *Mol. Endocrinol.* 17, 1144–1154. doi: 10.1210/me.2002-2366

Hong, M. J., Yoo, S. S., Choi, J. E., Kang, H. G., Do, S. K., Lee, J. H., et al. (2018). Functional intronic variant of SLC5A10 affects DRG2 expression and survival outcomes of early-stage non-small-cell lung cancer. *Cancer Sci.* 109, 3902–3909. doi: 10.1111/cas.13814

Huang, Z., Chen, X., and Chen, D. (2011). Myostatin: a novel insight into its role in metabolism, signal pathways, and expression regulation. *Cell. Signal.* 23, 1441–1446. doi: 10.1016/j.cellsig.2011.05.003

Ji, R., Cui, P., Ding, F., Geng, J., Gao, H., Zhang, H., et al. (2009). Monophyletic origin of domestic bactrian camel (Camelus bactrianus) and its evolutionary relationship with the extant wild camel (Camelus bactrianus ferus). *Anim. Genet.* 40, 377–382. doi: 10.1111/j.1365-2052.2008.01848.x

Ji, S., Losinski, R. L., Cornelius, S. G., Frank, G. R., Willis, G. M., Gerrard, D. E., et al. (1998). Myostatin expression in porcine tissues: tissue specificity and developmental and postnatal regulation. *Am. J. Physiol.* 275(4 Pt 2), R1265–R1273. doi: 10.1152/ajpregu.1998.275.4.R1265

Jiang, M. S., Liang, L. F., Wang, S., Ratovitski, T., Holmstrom, J., Barker, C., et al. (2004). Characterization and identification of the inhibitory domain of GDF-8 propeptide. *Biochem. Biophys. Res. Commun.* 315, 525–531. doi: 10.1016/j.bbrc.2004.01.085

Jones, S. (2004). An overview of the basic helix-loop-helix proteins. *Genome Biol.* 5:226. doi: 10.1186/gb-2004-5-6-226

Karlsson, A. H., Klont, R. E., and Fernandez, X. (1999). Skeletal muscle fibres as factors for pork quality. *Livest. Prod. Sci.* 60, 255–269. doi: 10.1016/S0301-6226(99)00098-6

Kim, J. S., Cross, J. M., and Bamman, M. M. (2005). Impact of resistance loading on myostatin expression and cell cycle regulation in young and older men and women. *Am. J. Physiol. Endocrinol. Metab.* 288, E1110–E1119. doi: 10.1152/ajpendo.00464.2004

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., et al. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* 6:e15925. doi: 10.1371/journal.pone.0015925

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Kugelberg, E. (1976). Adaptive transformation of rat soleus motor units during growth. *J. Neurol. Sci.* 27, 269–289. doi: 10.1016/0022-510x(76)90001-0

Lakshman, K. M., Bhasin, S., Corcoran, C., Collins-Racie, L. A., Tchistiakova, L., Forlow, S. B., et al. (2009). Measurement of myostatin concentrations in human serum: circulating concentrations in young and older men and effects

of testosterone administration. *Mol. Cell. Endocrinol.* 302, 26–32. doi: 10.1016/j.mce.2008.12.019

Lee, S. J. (2008). Genetic analysis of the role of proteolysis in the activation of latent myostatin. *PLoS One* 3:e1628. doi: 10.1371/journal.pone.0001628

Lee, S. J. (2010). Extracellular regulation of myostatin: a molecular rheostat for muscle mass. *Immunol. Endocr. Metab. Agents Med. Chem.* 10, 183–194. doi: 10.2174/187152210793663748

Lee, S. J., and McPherron, A. C. (2001). Regulation of myostatin activity and muscle growth. *Proc. Natl. Acad. Sci. U.S.A.* 98, 9306–9311. doi: 10.1073/pnas.151270098

Lee, S. Y., Hong, M. J., Jeon, H. S., Choi, Y. Y., Choi, J. E., Kang, H. G., et al. (2015). Functional intronic ERCC1 polymorphism from regulomeDB can predict survival in lung cancer after surgery. *Oncotarget* 6, 24522–24532. doi: 10.18632/oncotarget.4083

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Manabe, N., Myoumoto, A., Tajima, C., Fukumoto, M., Nakayama, M., Uchio, K., et al. (2000). Immunochemical characteristics of a novel cell death receptor and a decoy receptor on granulosa cells of porcine ovarian follicles. *Cytotechnology* 33, 189–201. doi: 10.1023/A:1008146119761

Matsakas, A., Bozzo, C., Cacciani, N., Caliaro, F., Reggiani, C., Mascarello, F., et al. (2006). Effect of swimming on myostatin expression in white and red gastrocnemius muscle and in cardiac muscle of rats. *Exp. Physiol.* 91, 983–994. doi: 10.1113/expphysiol.2006.033571

McPherron, A. C., Lawler, A. M., and Lee, S. J. (1997). Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. *Nature* 387, 83–90. doi: 10.1038/387083a0

Megeney, L. A., Kablar, B., Garrett, K., Anderson, J. E., and Rudnicki, M. A. (1996). MyoD is required for myogenic stem cell function in adult skeletal muscle. *Genes Dev.* 10, 1173–1183. doi: 10.1101/gad.10.10.1173

Mendes de Almeida, R., Tavares, J., Martins, S., Carvalho, T., Enguita, F. J., Brito, D., et al. (2017). Whole gene sequencing identifies deep-intronic variants with potential functional impact in patients with hypertrophic cardiomyopathy. *PLoS One* 12:e0182946. doi: 10.1371/journal.pone.0182946

Miura, T., Kishioka, Y., Wakamatsu, J., Hattori, A., Hennebry, A., Berry, C. J., et al. (2006). Decorin binds myostatin and modulates its activity to muscle cells. *Biochem. Biophys. Res. Commun.* 340, 675–680. doi: 10.1016/j.bbrc.2005.12.060

Montarras, D., Lindon, C., Pinset, C., and Domeyne, P. (2000). Cultured myf5 null and myoD null muscle precursor cells display distinct growth defects. *Biol. Cell.* 92, 565–572. doi: 10.1016/s0248-4900(00)01110-2

Morissette, M. R., Cook, S. A., Foo, S., McKoy, G., Ashida, N., Novikov, M., et al. (2006). Myostatin regulates cardiomyocyte growth through modulation of Akt signaling. *Circ. Res.* 99, 15–24. doi: 10.1161/01.RES.0000231290.45676.d4

Morrison, P. K., Bing, C., Harris, P. A., Maltin, C. A., Grove-White, D., and Argo, C. M. (2014). Post-mortem stability of RNA in skeletal muscle and adipose tissue and the tissue-specific expression of myostatin, perilipin and associated factors in the horse. *PLoS One* 9:e100810. doi: 10.1371/journal.pone.0100810

Mou, Z., Hyde, T. M., Lipska, B. K., Martinowich, K., Wei, P., Ong, C. J., et al. (2015). Human Obesity Associated with an Intronic SNP in the Brain-Derived Neurotrophic Factor Locus. *Cell. Rep.* 13, 1073–1080. doi: 10.1016/j.celrep.2015.09.065

Muzzachi, S., Oulmouden, A., Cherifi, Y., Yahyaoui, H., Zayed, M., Burger, P., et al. (2015). Sequence and polymorphism analysis of the camel (*Camelus dromedarius*) myostatin gene. *Emir. J. Food Agric.* 27, 367–373. doi: 10.9755/ejfa.v27i4.19910

Ostrovsky, O., Grushchenko-Polaq, A. H., Beider, K., Mayorov, M., Canaani, J., Shimoni, A., et al. (2018). Identification of strong intron enhancer in the heparanase gene: effect of functional rs4693608 variant on HPSE enhancer activity in hematological and solid malignancies. *Oncogenesis* 7:51. doi: 10.1038/s41389-018-0060-68

Pierri, C. L., Parisi, G., and Porcelli, V. (2010). Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. *Biochim. Biophys. Acta* 1804, 1695–1712. doi: 10.1016/j.bbapap.2010.04.008

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., et al. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics* 8:14. doi: 10.1186/1479-7364-8-14

Pirruccello-Straub, M., Jackson, J., Wawersik, S., Webster, M. T., Salta, L., Long, K., et al. (2018). Blocking extracellular activation of myostatin as a strategy for treating muscle wasting. *Sci. Rep.* 8:2292. doi: 10.1038/s41598-018-20524-20529

Reardon, K. A., Davis, J., Kapsa, R. M., Choong, P., and Byrne, E. (2001). Myostatin, insulin-like growth factor-1, and leukemia inhibitory factor mRNAs are upregulated in chronic human disuse muscle atrophy. *Muscle Nerve* 24, 893–899. doi: 10.1002/mus.1086

Rebbapragada, A., Benchabane, H., Wrana, J. L., Celeste, A. J., and Attisano, L. (2003). Myostatin signals through a transforming growth factor beta-like signaling pathway to block adipogenesis. *Mol. Cell. Biol.* 23, 7230–7242. doi: 10.1128/mcb.23.20.7230-7242.2003

Rodriguez, J., Vernus, B., Chelh, I., Cassar-Malek, I., Gabillard, J. C., Hadj Sassi, A., et al. (2014). Myostatin and the skeletal muscle atrophy and hypertrophy signaling pathways. *Cell. Mol. Life Sci.* 71, 4361–4371. doi: 10.1007/s00018-014-1689-x

Sailsbery, J. K., and Dean, R. A. (2012). Accurate discrimination of bHLH domains in plants, animals, and fungi using biologically meaningful sites. *BMC Evol. Biol.* 12:154. doi: 10.1186/1471-2148-12-154

Santiago, C., Ruiz, J. R., Rodriguez-Romo, G., Fiuza-Luces, C., Yvert, T., Gonzalez-Freire, M., et al. (2011). The K153R polymorphism in the myostatin gene and muscle power phenotypes in young, non-athletic men. *PLoS One* 6:e16323. doi: 10.1371/journal.pone.0016323

Sazili, A. Q., Parr, T., Sensky, P. L., Jones, S. W., Bardsley, R. G., and Buttery, P. J. (2005). The relationship between slow and fast myosin heavy chain content, calpastatin and meat tenderness in different ovine skeletal muscles. *Meat Sci.* 69, 17–25. doi: 10.1016/j.meatsci.2004.06.021

Schuelke, M., Wagner, K. R., Stolz, L. E., Hubner, C., Riebel, T., Komen, W., et al. (2004). Myostatin mutation associated with gross muscle hypertrophy in a child. *N. Engl. J. Med.* 350, 2682–2688. doi: 10.1056/NEJMoa040933

Shah, M. G., Qureshi, M., Reissmann, A. S., and Schwartz, H. J. (2006). Sequencing and sequence analysis of myostatin gene in the exon 1 of the Camel (*Camelus dromedarius*). *Pak. Vet. J.* 26, 176–178.

Sharma, M., Kambadur, R., Matthews, K. G., Somers, W. G., Devlin, G. P., Conaglen, J. V., et al. (1999). Myostatin, a transforming growth factor-beta superfamily member, is expressed in heart muscle and is upregulated in cardiomyocytes after infarct. *J. Cell. Physiol.* 180, 1–9. doi: 10.1002/(sici)1097-4652(199907)180:1<1::aid-jcp1>3.3.co;2-m

Shyu, K. G., Lu, M. J., Wang, B. W., Sun, H. Y., and Chang, H. (2006). Myostatin expression in ventricular myocardium in a rat model of volume-overload heart failure. *Eur. J. Clin. Invest.* 36, 713–719. doi: 10.1111/j.1365-2362.2006.01718.x

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. doi: 10.1038/msb.2011.75

Spiller, M. P., Kambadur, R., Jeanplong, F., Thomas, M., Martyn, J. K., Bass, J. J., et al. (2002). The myostatin gene is a downstream target gene of basic helix-loop-helix transcription factor MyoD. *Mol. Cell. Biol.* 22, 7066–7082. doi: 10.1128/mcb.22.20.7066-7082.2002

Szlama, G., Trexler, M., Buday, L., and Patthy, L. (2015). K153R polymorphism in myostatin gene increases the rate of promyostatin activation by furin. *FEBS Lett.* 589, 295–301. doi: 10.1016/j.febslet.2014.12.011

Szlama, G., Trexler, M., and Patthy, L. (2013). Latent myostatin has significant activity and this activity is controlled more efficiently by WFIKKN1 than by WFIKKN2. *FEBS J.* 280, 3822–3839. doi: 10.1111/febs.12377

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tellgren, A., Berglund, A. C., Savolainen, P., Janis, C. M., and Liberles, D. A. (2004). Myostatin rapid sequence evolution in ruminants predates domestication. *Mol. Phylogenet. Evol.* 33, 782–790. doi: 10.1016/j.ympev.2004.07.004

Thies, R. S., Chen, T., Davies, M. V., Tomkinson, K. N., Pearson, A. A., Shakey, Q. A., et al. (2001). GDF-8 propeptide binds to GDF-8 and antagonizes biological activity by inhibiting GDF-8 receptor binding. *Growth Factors* 18, 251–259. doi: 10.3109/08977190109029114

Torrella, J. R., Whitmore, J. M., Casas, M., Fouces, V., and Viscor, G. (2000). Capillarity, fibre types and fibre morphometry in different sampling sites across and along the tibialis anterior muscle of the rat. *Cells Tissues Organs* 167, 153–162. doi: 10.1159/000016778

Tsunoda, T., and Takagi, T. (1999). Estimating transcription factor bindability on DNA. *Bioinformatics* 15, 622–630. doi: 10.1093/bioinformatics/15.7.622

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43 11, 11–33. doi: 10.1002/0471250953.bi1110s43

Vaz-Drago, R., Custodio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Hum Genet* 136, 1093–1111. doi: 10.1007/s00439-017-1809-4

Walker, R. G., Angerman, E. B., Kattamuri, C., Lee, Y. S., Lee, S. J., and Thompson, T. B. (2015). Alternative binding modes identified for growth and differentiation factor-associated serum protein (GASP) family antagonism of myostatin. *J. Biol. Chem.* 290, 7506–7516. doi: 10.1074/jbc.M114.624130

Walton, K. L., Makanji, Y., Chen, J., Wilce, M. C., Chan, K. L., Robertson, D. M., et al. (2010). Two distinct regions of latency-associated peptide coordinate stability of the latent transforming growth factor-beta1 complex. *J. Biol. Chem.* 285, 17029–17037. doi: 10.1074/jbc.M110.110288

Wehling, M., Cai, B., and Tidball, J. G. (2000). Modulation of myostatin expression during modified muscle use. *FASEB J.* 14, 103–110. doi: 10.1096/fasebj.14.1.103

Wolfman, N. M., McPherron, A. C., Pappano, W. N., Davies, M. V., Song, K., Tomkinson, K. N., et al. (2003). Activation of latent myostatin by the BMP-1/tolloid family of metalloproteinases. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15842–15846. doi: 10.1073/pnas.2534946100

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188. doi: 10.1038/ncomms6188

Xie, Y., Perry, B. D., Espinoza, D., Zhang, P., and Price, S. R. (2018). Glucocorticoid-induced CREB activation and myostatin expression in C2C12 myotubes involves phosphodiesterase-3/4 signaling. *Biochem. Biophys. Res. Commun.* 503, 1409–1414. doi: 10.1016/j.bbrc.2018.07.056

Zimmers, T. A., Davies, M. V., Koniaris, L. G., Haynes, P., Esquela, A. F., Tomkinson, K. N., et al. (2002). Induction of cachexia in mice by systemically administered myostatin. *Science* 296, 1486–1488. doi: 10.1126/science.1069525

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer RM declared a shared affiliation, with no collaboration, with several of the authors MF, LG, CP, and EC to the handling Editor at the time of review.

Copyright © 2019 Favia, Fitak, Guerra, Pierri, Faye, Oulmouden, Burger and Ciani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 June 2019
doi: 10.3389/fgene.2019.00568

TYR Gene in Llamas: Polymorphisms and Expression Study in Different Color Phenotypes

*Melina Anello[1], Estefanía Fernández[1], María Silvana Daverio[1,2], Lidia Vidal-Rioja[1] and Florencia Di Rocco[1]**

[1] *Laboratorio de Genética Molecular, Instituto Multidisciplinario de Biología Celular, CONICET–UNLP–CIC, La Plata, Argentina,* [2] *Cátedra de Biología, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina*

Tyrosinase, encoded by *TYR* gene, is an enzyme that plays a major role in mammalian pigmentation. It catalyzes the oxidation of L-dihydroxy-phenylalanine (DOPA) to DOPA quinone, a precursor of both types of melanin: eumelanin and pheomelanin. *TYR* is commonly known as the *albino* locus since mutations in this gene result in albinism in several species. However, many other *TYR* mutations have been found to cause diluted phenotypes, like the Himalayan or chinchilla phenotypes in mice. The llama (*Lama glama*) presents a wide variety of coat colors ranging from non-diluted phenotypes (eumelanic and pheomelanic), through different degrees of dilution, to white. To investigate the possible contribution of *TYR* gene to coat color variation in llamas, we sequenced *TYR* exons and their flanking regions and genotyped animals with diluted, non-diluted, and white coat, including three blue-eyed white individuals. Moreover, we analyzed mRNA expression levels in skin biopsies by qPCR. *TYR* coding region presented nine SNPs, of which three were non-synonymous, c.428A > G, c.859G > T, and c.1490G > T. We also identified seven polymorphisms in non-coding regions, including two microsatellites, an homopolymeric repeat, and five SNPs: one in the promoter region (c.1-26C > T), two in the 3′-UTR, and two flanking the exons. Although no complete association was found between coat color and SNPs, c.1-26C > T was partially associated to diluted phenotypes. Additionally, the frequency of the G allele from c.428A > G was significantly higher in white compared to non-diluted. Results from qPCR showed that expression levels of TYR in white llamas were significantly lower ($p < 0.05$) than those in diluted and non-diluted phenotypes. Screening for variation in regulatory regions of TYR did not reveal polymorphisms that explain such differences. However, data from this study showed that TYR expression levels play a role in llama pigmentation.

Keywords: tyrosinase, coat color, dilution, *Lama glama*, polymorphisms, expression

OPEN ACCESS

Edited by:
Pamela Burger, University of Veterinary Medicine, Vienna, Austria

Reviewed by:
Susana Seixas, University of Porto, Portugal
Abdussamad Muhammad Abdussamad, Bayero University Kano, Nigeria

****Correspondence:***
Florencia Di Rocco
fdirocco@imbice.gov.ar

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *29 October 2018*
Accepted: *29 May 2019*
Published: *12 June 2019*

Citation:
Anello M, Fernández E, Daverio MS, Vidal-Rioja L and Di Rocco F (2019) TYR Gene in Llamas: Polymorphisms and Expression Study in Different Color Phenotypes. Front. Genet. 10:568. doi: 10.3389/fgene.2019.00568

INTRODUCTION

In mammals, basic coat colors are defined by the relative proportion between two types of melanin: eumelanin (black or brown) and pheomelanin (yellow or red). At molecular level, eumelanin:pheomelanin ratio is regulated mainly by the ligand–receptor system of the agouti signaling protein (*ASIP*) and the melanocortin 1-receptor (*MC1R*). Additionally, color phenotype

depends on the expression and interaction of many other genes that can disrupt the normal pigmentation pathway (Cieslak et al., 2011).

Tyrosinase, encoded by *TYR* gene, is a key enzyme for melanin synthesis. This copper-containing enzyme catalyzes the first two steps in the melanin biosynthesis pathway, converting tyrosine to L-dihydroxy-phenylalanine (DOPA) and afterward DOPA to DOPA quinone, a precursor of both types of melanin. *TYR* is commonly known as the albino locus since mutations in this gene result in albinism in several species, including humans. More than 100 mutations in the *TYR* gene have been identified in people with oculocutaneous albinism type 1 (The Albinism Database[1]). Because of the lack of melanin production, these patients present white hair, light-colored eyes, and a very pale skin that does not tan. Albinism has also been described in mice (Yokoyama et al., 1990; Beermann et al., 2004), cats (Imes et al., 2006), cattle (Schmutz et al., 2004), rabbits (Aigner et al., 2000), buffalos (Damé et al., 2012), donkeys (Utzeri et al., 2016) among other species.

Additionally, other *TYR* mutations have been found to cause milder phenotypes. For example, the Himalayan phenotype implicates different mutations that result in a temperature-regulated activity of *TYR*, where cooler parts of the body are pigmented while warmer parts remain white. These mutations have been described in mice (Kwon et al., 1989), minks (Benkel et al., 2009), rabbits (Aigner et al., 2000), and cats (Lyons et al., 2005). Another example is the chinchilla allele from mice, which encodes a tyrosinase whose activity is from one-third to one-half that of the normal. This is caused by a point mutation in the *TYR* gene and chinchilla mice exhibit a grayish color (Lamoreux et al., 2001). Another point mutation in *TYR* gene is responsible for the mice platinum phenotype, which results in animals with almost complete loss of pigmentation (Orlow et al., 1993). Recently, two near-white *TYR* alleles were described in mice, Dhoosara and Chandana, characterized by a marked hypopigmentation in the body and the eyes (Challa et al., 2016).

Most of what is known about the regulation of *TYR* gene derives from studies carried out in mice and extended to human (Giraldo et al., 2003; Murisier and Beermann, 2006; Reinisalo et al., 2012). According to these studies, *TYR* is regulated by a combination of proximal promoter elements and far upstream regulatory sequences, being the locus control region (LCR) the most important one.

The llama (*Lama glama*) is a South American camelid which presents a great diversity of coat colors, including black, chocolate brown, many shades of light brown (from red to pale cream), and complete white coat (Frank et al., 2006). Occasionally, some white llamas present blue eyes. Patterns and spotted phenotypes are also frequent in llamas.

The molecular mechanisms that control pigmentation in camelids have mainly been studied in alpacas (*Vicugna pacos*). Different authors have studied *MC1R* (Powell et al., 2008; Feeley and Munyard, 2009; Guridi et al., 2011; Chandramohan et al., 2015), *ASIP* (Feeley et al., 2011; Chandramohan et al., 2013), *TYR*, and *MATP* genes (Cransberg and Munyard, 2011) in relation to alpaca coat colors. However, in llamas the mechanisms that control pigmentation are not fully understood. In a previous work, we analyzed *MC1R* and *ASIP* genes and found two *ASIP* polymorphisms associated with eumelanic phenotype (Daverio et al., 2016). We also studied *KIT* and *MITF-M* genes and their relation to white coat. Although no variants were found to be associated with white phenotype, both genes were less expressed in this phenotype (Anello et al., 2019). Here, we aimed to study the possible contribution of *TYR* gene to coat color variation in llamas and to do so, we describe the *TYR* gene, its variation, and its skin expression.

MATERIALS AND METHODS

Samples

All samples were collected following the recommendations of the Argentine Ethical Guidelines for Biomedical Investigation in Animals from Laboratory, Farm, or Obtained from Nature (Resolution No. 1047/05 from CONICET, Argentina). The protocol was approved by the Institutional Committee for the Care and Use of Laboratory Animals (CICUAL) from the Multidisciplinary Institute of Cellular Biology (IMBICE). All llama owners provided an informed consent for their animal's inclusion prior to sampling and were present during the collection of the samples.

To determine *TYR* sequence and variation in llamas, blood samples were collected from 29 unrelated animals from different breeding farms in Argentina. Eighty-five additional samples were included for genotyping of polymorphisms of interest. All samples were taken by jugular vein puncture from young-adult animals. Approximately equal sex proportions were sampled and the lack of relationship between the sampled animals was verified by consulting breeder records. Coat color phenotype for each sample was documented and pictures were taken whenever possible.

Phenotypes sampled were divided into three groups for the subsequent analyses: NON-DILUTED (n = 47), DILUTED (n = 27), and WHITE (n = 40). The first group included the following phenotypes: BLACK, eumelanic llamas with non-diluted black or dark brown pigmented coats; RED, llamas with pheomelanic red coat, completely pigmented; and BLACK FACE, animals that presented non-diluted pheomelanic pigmentation with black face and legs, a very common phenotype among llamas. On the contrary, DILUTED PHENOTYPES showed either a eumelanin or pheomelanin dilution, so it included GRAY, individuals with diluted eumelanic coats, and FAWN, llamas with diluted pheomelanin (light brown, fawn, and cream coats). Finally, WHITE included two phenotypes, WHITE, non-albino llamas with dark eyes and a full white coat; and BLUE-EYED WHITE, individuals with full white coat and blue eyes. **Figure 1** shows some examples of llama phenotypes included in this study. Pictures of the blue-eyed white phenotype are not available. More information about samples used in this paper can be found in **Supplementary Table 1**.

For the expression analysis of *TYR*, skin biopsies from pheomelanic NON-DILUTED, pheomelanic DILUTED, and

[1] http://www.ifpcs.org/albinism/oca1mut.html

FIGURE 1 | Photographs of llamas illustrating color phenotypes included in this study. **(A)** Black, **(B)** Red, **(C)** Black Face, **(D)** Gray, **(E)** Fawn, and **(F)** White.

WHITE llamas (three animals of each phenotype) were collected using a disposable 3-mm diameter biopsy punch. To avoid RNA degradation, samples were kept in RNAlater (SIGMA, Germany) during their transportation to the laboratory, where they were immediately processed.

DNA Extraction and PCR Amplification

Total genomic DNA was extracted from blood samples using standard procedures (Sambrook and Russell, 2001). DNA was resuspended in TE buffer and stored at −20°C for further analysis.

Primers flanking each of the five exons were designed over the alpaca genome available at Genome Browser (GCA_000164845.2). **Supplementary Table 2** shows primers sequences and amplicon length.

Additionally, two regions of *TYR* promoter were amplified in 15 samples. A 792-bp fragment corresponding to *TYR* proximal protomer region and a 789-bp fragment where the LCR was located, at approximately at −9 kb. Since information about this latter region was restricted to mouse and human (AF364302.1 and AY180962.1) we aligned them with genomes scaffolds from other camelids available at GenBank (*Camelus dromedarius* NW_011591148.1, *Camelus bactrianus* NW_011544909.1, *Camelus ferus* NW_006211950.1, and *V. pacos* NW_005883058.1) in order to designed primers (**Supplementary Table 2**).

Polymerase chain reaction amplification reactions were performed with Taq DNA polymerase (PB-L, Pegasus) and 1× PCR buffer (PB-L, Pegasus) according to manufacturer's instructions. The cycling profile consisted of an initial denaturation step at 94°C for 3 min, 30 cycles of 40 s at 94°C, 50 s at 58–65°C, 40 s at 72°C, and a 5-min final extension at 72°C. PCR products were checked on a 1% agarose gel stained with GelRed™, purified by PEG precipitation, and sequenced by Macrogen Inc., Korea.

The sequences obtained were aligned and analyzed to identify polymorphisms using Geneious software (v.6.1.8, Biomatters). Elements in the regulatory regions were identified by alignment with mouse and human sequences.

Protein topology and domains were predicted using Constrained Consensus TOPology prediction server (CCTOP) (Dobson et al., 2015) and Conserved Domains Database (CDD) (Marchler-Bauer et al., 2017) software, respectively, and compared with UniProt reviewed entries (Bateman et al., 2017). TMHMM server[2] was used for the prediction of transmembrane regions. Llama TYR protein was aligned to the homologous proteins of the following species: *V. pacos* (XP_006218431.1), *Ovis aries* (NP_001123499.1), *Capra hircus* (NP_001274491.1), *Homo sapiens* (AAB60319.1), *Mus musculus* (AAA40516.1), *Desmodus rotundus* (XP_024428480.1), *Canis lupus familiaris* (NP_001002941), *Delphinapterus leucas* (XP_022449451), and *Neovison vison* (AJO15925.1).

Genotyping and Analysis of Polymorphisms of Interest

Single-nucleotide polymorphisms c.1-26C > T and c.1490G > T were genotyped by allele-specific PCR amplification. For each variant, a pair of primers was designed with the corresponding complementary nucleotide in the 3′-end of the forward primer (**Supplementary Table 2**). Additionally, a destabilizing mismatch was introduced within the 3′-end of the allele-specific primers to improve specificity. Each reaction was carried out in a separate tube and allele products, which presented different lengths, were further identified by 2% agarose gel electrophoresis (for 90 min at 90 V) stained with GelRed™. Positive sequenced controls for each variant determination were used.

The SNPs c.428A > G and c.859G > T were genotyped by PCR–RFLP using sequenced samples as controls. PCR products from exon 1 were digested with the enzyme *Rsa1* that recognizes the c.428A > G mutation. If the amplicon presented allele A, the enzyme cut it into two fragments of 818 and 140 bp while if the amplicon presented the G allele, a new restriction site was created and digestion resulted in three fragments of 506, 312, and 140 bp. Digestion mix consisted of 7 μl of PCR product, 5 U of *Rsa1* (Thermo Fisher Scientific Inc.), and 1.5 μl of 10× Buffer Tango. Digestions were incubated at 37°C overnight and restriction fragments were analyzed by electrophoresis on a 2% agarose gel

[2]http://www.cbs.dtu.dk/services/TMHMM

for 90 min at 90 V, stained with GelRed™. PCR products from exon 2 were digested with the enzyme *BsmAI* to genotype the SNP c.859G > T. If the product had the allele T, *BsmAI* cut it into two fragments of 433 and 88 bp while if it presented the G allele, it remained undigested. Digestion mix consisted of 7 μl of PCR product, 5 U of *BsmAI* (New England BioLabs Inc.), and 1.5 μl of 10× Buffer 3. Digestions were incubated at 55°C overnight and restriction fragments were analyzed by 8% acrylamide gel electrophoresis (180 min at 160 V) stained with GelRed™.

Association between allelic frequencies or genotypes and phenotypes was determined by Fisher's exact test or Monte Carlo method, depending on the data set distribution, using PASW Statistics 1.8 (SPSS Inc., 2009). Association between haplotypes for the three non-synonymous SNPs was performed using the same methods, including samples whose haplotypes could be determined without ambiguities.

Web servers PolyPhen-2[3], SIFT[4], and PROVEAN Protein[5] were used to predict the possible impact of amino acid substitutions on TYR protein. PolyPhen-2 uses a sequence- and structure-based approach and classifies the SNPs as benign, possibly damaging, or probably damaging using a position-specific independent count score which ranges from 0 to 1. SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids and classifies substitutions in tolerant or intolerant based on scores (damaging if the score is =0.05 and tolerated if the score is >0.05). PROVEAN Protein applies an evolutionary conservation-based method and scores the SNP as neutral or deleterious (numbers equal or below −2.5 are considered deleterious).

RNA Extraction and Expression Analysis

Total RNA was extracted from skin biopsies by homogenization in TRIzol® according to manufacturer's instructions. Reverse transcription was performed to obtain cDNA, using RevertAid Reverse Transcriptase (Thermo Fisher Scientific Inc.), and random primers (Biodynamics), following the manufacturer's instructions.

Quantitative real-time PCR was carried out using a pair of primers specifically designed for this assay, where one primer annealed over exon–exon junction, avoiding genomic amplification (**Table 1**). Ribosomal *18S*, which has been previously used in melanogenesis expression studies (Saravanaperumal et al., 2014), was used for data normalization.

Amplification reactions were carried out in a RotorGene Q (Qiagen) and consisted of 20 μl, including 4 HOT FIREPol® EvaGreen® qPCR Mix Plus (ROX) (Solis Biodyne), 0.5 mM of each primer, and 1 ng cDNA. The cycling parameters were: 15 min at 95°C, 40 cycles of 15 s at 95°C, 20 s at 60°C, 20 s at 72°C, and a final gradient from 95 to 72°C. qPCR reaction were optimized; PCR efficiencies calculated from the slope were within 90–110%, r^2 over 98%. Each gene was amplified in three technical replicates for every sample and two NTC controls. Melting curve analysis was performed following amplification to verify the absence of non-specific amplification or primer dimer. Ct was determined by Rotor-Gene Q Pure Detection software version 2.3.1. Quantification of transcript abundance was carried out using the comparative threshold cycle (Ct) method by Livak and Schmittgen (2001), and ANOVA was used to assess if differences in expression were significant.

[3]http://genetics.bwh.harvard.edu/pph2/

[4]https://sift.bii.a-star.edu.sg/

[5]http://provean.jcvi.org/seq_submit.php

RESULTS

Description of *TYR* Gene

We sequenced the *TYR* five exons with its flanking regions, including 58 bp from the 5′-untranslated region (UTR) and 85 bp from the 3′-UTR. The entire coding sequence comprised 1593 bp, divided into: 819 bp for exon 1, 218 bp for exon 2, 147 bp for exon 3, 182 bp for exon 4, and 227 bp for exon 5. Llama *TYR* gene sequence was deposited in GenBank under the accession number MK089778.

The protein encoded by the *TYR* gene is predicted to have 530 amino acids. It presents a signal peptide of 18 residues, an EGF-like domain (from amino acid 57 to 113) and a transmembrane region (from 474 to 496). The characteristic tyrosinase domain that binds two copper ions, via two sets of three histidine, expands from amino acid 170 to 403 (**Supplementary Figure 1**). Identity to orthologous proteins was as high as 99% with alpaca, 91% with sheep and goat, 88% with human, and 85% with mouse.

TYR Polymorphisms and Coat Color Phenotype

To study the variation of *TYR* gene, the exons and their flanking regions were sequenced (3058 bp in total) in 29 llamas from different origins and with different color phenotypes. A total of 17 polymorphisms were detected. **Figure 2** shows the distribution of polymorphisms along the gene. The coding region showed nine SNPs, three of which were non-synonymous: c.428A > G, located in exon 1, produces an amino acid change in the protein from histidine to arginine in residue 143 of the enzyme; c.859G > T was detected in exon 2 and it changes the alanine 287 to serine; and c.1490G > T, in exon 5, replaces the arginine in position 497 to leucine. Additionally, a C/T transition was observed 26 bp before the initial codon ATG and two consecutive SNPs were detected in the 3′-UTR region, 14 and 15 bp after the stop codon (c.1593+14T > C and c.1593+15G > T). Six synonymous SNPs were observed, in exons 1 and 5.

Furthermore, variation in the intronic regions was detected: two SNPs, two microsatellites, and one homopolymeric repeat were identified in the flanking regions of *TYR* exons (**Figure 2**). The first microsatellite was in intron 2, 27 bp after the end of exon 2, and consisted of five nucleotides (TTTCC) that repeated a variable number of times. Alleles variated in the number of repeats, between 20 and 30, and presented imperfections in their motif. According to the motif, we classified the alleles into three types: type 1 consisted of the five nucleotides TTTCC repeated, while type 2 added the pattern (TT)-(TTTCC)2-(TT) in between the repeats. Type 3 presented a 1 bp deletion at the beginning (TTTC) and a 1 bp substitution in some other repeats (TCTCC).

TABLE 1 | Distribution of genotypes among phenotypic groups.

		c.1-26C/T			c.428A > G			c.859G > T			c.1490G > T		
		C/C	C/T	T/T	A/A	A/G	G/G	G/G	G/T	T/T	G/G	G/T	T/T
NON-DILUTED PHENOTYPES	BLACK	20	3	–	22	1	–	13	4	6	17	6	1
	RED	7	2	1	9	1	–	8	1	1	8	1	1
	BLACK FACE	9	2	2	12	1	–	13	–	–	5	7	1
	TOTAL	36	7	3	43	3	–	34	5	7	30	14	3
DILUTED PHENOTYPES	FAWN	11	7	1	19	1	–	13	1	3	9	8	3
	GRAY	2	3	–	6	1	–	5	–	–	3	4	–
	TOTAL	13	10	1	25	2	–	18	1	3	12	12	3
WHITE PHENOTYPES	WHITE	33	4	–	29	6	2	26	5	6	26	10	1
	BLUE-EYED WHITE	1	2	–	2	1	–	3	–	–	1	2	–
	TOTAL	34	6	0	31	7	2	29	5	6	27	12	1

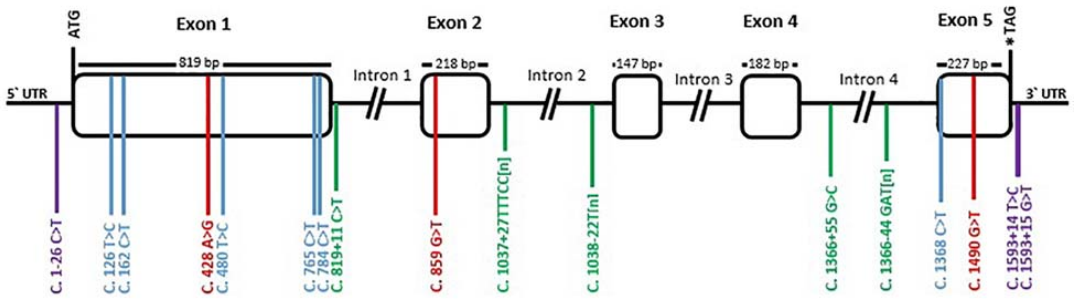


FIGURE 2 | Schematic representation of *TYR* gene with its polymorphism's distribution in llamas. Rectangular boxes correspond to exons while introns and non-translated regions are represented by black lines. Positions +1 was assigned to the translation initiation site (ATG codon) and * indicates the stop codon. Polymorphisms numbering is relative to the coding DNA sequence and indicated by transversal colored lines. Violet correspond to mutations in the UTR regions, blue correspond to synonymous mutation in the coding region, red to non-synonymous mutations, and green to intronic mutations.

This last allele was observed to segregate together with the T/T variant of SNP c.859G > T. An example of each type and a list of the observed alleles are provided in **Supplementary Figure 2** and **Supplementary Table 3**. However, it is most likely that there exist more alleles, since there were heterozygous animals for which was not possible to determine the alleles by direct sequencing.

The second microsatellite was detected in intron 4, 44 bp before the start of exon 5 and it consisted of three nucleotides (GAT) that repeated 13–15 times. Occasionally, in some samples, one repeat of the motif GAT was replaced by AAT. The list presented in **Supplementary Table 4** shows the observed genotypes and the possible combination of alleles.

Lastly, an homopolymeric repeat of thymidine was observed 22 bp before the start of exon 3. It presented variable length of 9 or 10 consecutive thymidine.

From all the variation found in *TYR* gene, we analyzed the polymorphisms that might influence coat color phenotype in 85 additional llama samples, that is the non-synonymous SNPs and the one located 26 bp upstream the ATG codon. Distribution of each genotype per color group is shown in **Table 1**. Regarding the coding SNPs, no significant association between any of the genotypes and color groups was found. However, distribution of allelic frequencies for the SNP c.428A > G was significantly different among groups (p = 0.015). The G allele was more frequent in the white group compared to non-diluted (p = 0.026) but the difference was not statistically different when compared to diluted (p = 0.103).

Furthermore, analysis of haplotypes for the three non-synonymous SNP did not show association with coat color ($p > 0.05$) (**Supplementary Table 5**).

Results from PolyPhen-2, SIFT, and PROVEAN Protein classified the three SNPs as benign/neutral/tolerated. The SNP c.428A > G deprives the protein of histidine 143, replacing it with another basic amino acid, arginine. Although this position is relatively conserved among species, bats also present an arginine residue (**Supplementary Figure 1**). The substitution c.859G > T encodes a change from alanine to serine in residue 287, a fairly common substitution (**Supplementary Figure 1**). Lastly, the SNP c.1490G > T changes arginine to leucine at amino acid 497, which is located at the end of the transmembrane region. This same mutation was reported in the alpaca *TYR* gene, and it was predicted to have a trivial biological effect (Cransberg and Munyard, 2011).

For the non-coding SNP c.1-26C/T, distribution among phenotypes resulted significantly different (p = 0.023). The heterozygous C/T genotype was significantly more frequent in the diluted group compared to non-diluted (p = 0.042) and white (0.013) (**Table 1**).

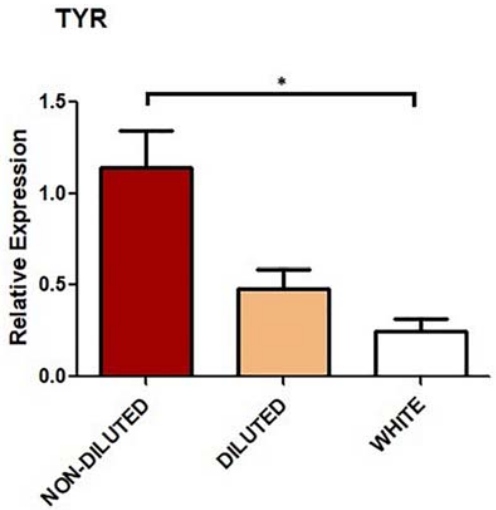


FIGURE 3 | Relative expression of *TYR* in the skin of llamas. Bar graphs show mean values. Error bars represent standard error (*SE*). Three biological replicates were analyzed for each color group. Asterisk indicates statistically significant differences ($p < 0.05$).

Expression of Llama *TYR* in Different Color Phenotypes

Housekeeping gene *18S* and *TYR* amplicons were observed in every sample analyzed and melting curves confirmed a single product in each PCR. NTC controls showed no amplification. Expression levels of *TYR* for NON-DILUTED, DILUTED, and WHITE phenotypes were compared. Results showed that the expression level in the WHITE group was significantly lower ($p < 0.05$) than in the NON-DILUTED group. Although expression level was lower in WHITE llamas compared to DILUTED, this difference was not significant. Neither was the difference in expression between NON-DILUTED and DILUTED (**Figure 3**).

Regulatory Regions of *TYR* Gene in Llamas

TYR proximal promoter (GenBank accession number MK847855) showed three conserved elements, two E-boxes (CATGTG) and one M-box (AGTCATGTGCT), which are known binding sites for bHLH-LZ transcription factors, like MITF. It also presented two binding sites for the orthodenticle homeobox 2 (OTX2), which is another transcription factor that is not involved in melanogenesis, but it is important in the retinal pigment epithelium. Additionally, a sequence corresponding to the TATA box was identified between 110 and 104 bp upstream of the first codon. No variation was observed within these regulatory elements (**Figure 4A**).

Besides the SNP c.1-26C > T, already described, another C to T SNP was detected at −402 bp, where the three possible genotypes were observed in the different phenotypic groups. Additionally, two other polymorphisms were detected in less frequency: an homopolymeric repeat of nine thymidine, that was found to be heterozygote for a T deletion in three llamas, one of each phenotypic group; and an imperfect repeated motif of ATT, that commonly presented 11 repetitions except for two individuals, one diluted and one non-diluted, that presented one repeat more in heterozygosity.

Figure 4B shows the LCR region (GenBank accession number MK847856). The position of Boxes A and B, corresponding to the putative binding sites for transcription factors, are indicated and aligned to the ones of mouse and human (X76647 and AY180962). Polymorphisms found in this region, SNPs and a 1 bp deletion, are also marked in **Figure 4B**. All of them were located outside Boxes A and B except for a T to C transition, within Box A. However, that position is not conserved between human and mouse, since one presents a C and the other a T, as it was observed in llamas. As for the other polymorphisms, different genotypes were observed among the different color phenotypes.

DISCUSSION

The structure of the llama *TYR* gene was similar to that previously described in human and other species (García-Borrón and Solano, 2002; Kanteev et al., 2015); the exons were relatively conserved and so were the domains of the encoded protein. Furthermore, identity values compared to orthologous proteins were high.

In reference to genetic variation, llamas presented several polymorphisms along *TYR* gene that included different types of variation. Nevertheless, exons 3 and 4, which codify the active region of the enzyme, did not present any mutations.

Different repeated sequences were observed in non-coding regions of *TYR* gene: microsatellites and homopolymeric repeats. Although it is not completely understood yet, there is strong evidence that repeated elements do have a role in the genome and that their distribution is not random (Vieira et al., 2016). There are numerous studies indicating that repeats located in introns can affect gene transcription, mRNA splicing, or export to cytoplasm (Li et al., 2004). Moreover, some of these repeats can have an effect on the phenotype and the length of microsatellites is particularly important in this aspect (Sjakste et al., 2013). Long alleles are associated to high predisposition to pathologies (Belguith-Maalej et al., 2013). In relation to coat color, a recent study revealed the presence of a large intronic insertion in Tyrosinase Associated Protein 1 (*TYRP1*) of the American Mink that alters the splicing of the gene and produces the American Palomino phenotype (Cirera et al., 2016). Due to the complexity of the repeated sequences observed in llamas, that variated in length and in motif, it was not possible to determine the total number of alleles. A more thorough study should be carried out to analyze these variants in particular and their relation with coat color. Nevertheless, we considered important to describe them as detailed as possible since there are few microsatellites described for camelids in comparison with other species and they have important applications for other studies such as linkage analysis, evolution, forensics, and population genetics.

Usually, disrupting mutations in *TYR* gene, like frameshifts or nonsense mutations, result in an inactive truncated protein

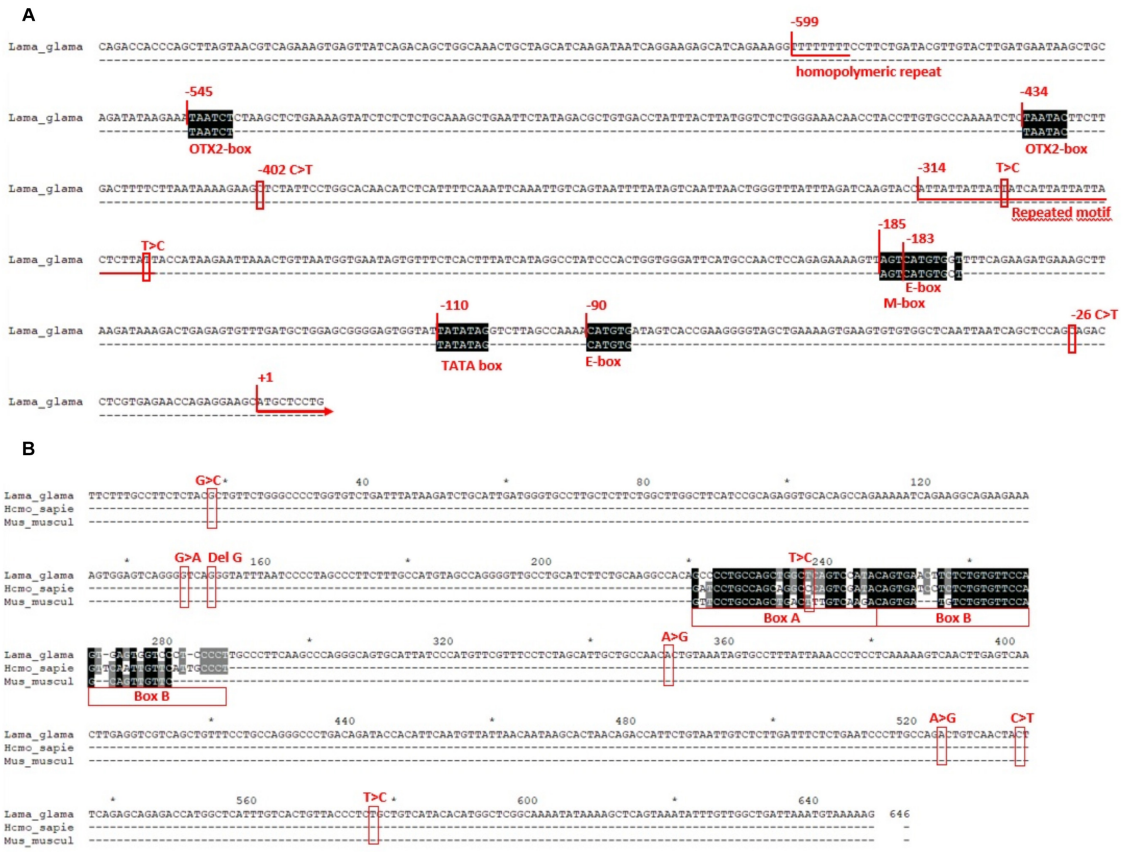


FIGURE 4 | (A) Proximal promoter region of *TYR* gene. Positions +1 were assigned to the translation initiation site (ATG codon). Regulatory elements are aligned with its consensus sequence in black blocks and named underneath. Numbers above the elements indicate relative positions. Red squares indicate the location of the SNPs found in llamas and the repeated elements are underlined. **(B)** TYR LCR region. Nucleotide positions (from 1 to 646) are indicated in black above the sequence. Number 1 was arbitrarily designed to the first nucleotide showed. Boxes A and B are aligned to the ones of mouse and human (X76647 and AY180962) and the variation found in llamas is squared in red color.

and have been reported to produce albino animals in different species (Oetting, 2000; Schmutz et al., 2004; Blaszczyk et al., 2005; Imes et al., 2006; Anistoroaei et al., 2008). Hypopigmentation, on the other side, has been more frequently associated with missense point mutations. Accordingly, there are at least seven alleles described that result in hypopigmentation in mice (Challa et al., 2016). Minks also present the Himalayan allele, that produces an hypopigmented phenotype and is the result of a single-nucleotide substitution (Benkel et al., 2009). So are the mutations associated with the diluted phenotypes Siamese and Burmese in cats (Lyons et al., 2005). We have analyzed TYR gene in llamas with non-diluted, diluted, and white phenotypes and three non-synonymous SNPs were observed. Although *in silico* analysis predicted the three SNPs to produce benign changes in the protein, the frequency of the G allele from c.428A > G was significantly higher in white compared to non-diluted, suggesting phenotype association. This allele produces an amino acid change from histidine to arginine in residue 143 of tyrosinase. Even though it is a relatively conserved position among species, the homologous protein from vampire bats (*D. rotundus*) presents an arginine in this position. Additionally, we searched the human data base from The Ensembl project[6] and found that the same mutation was observed with a very low frequency (0.00006153), but it was not reported if there is an associated phenotype. Tanking into account all the information just mentioned, it is difficult to assess the possible functional relevance of this variant.

Besides the study of polymorphisms, we focused on the expression of *TYR* gene. We found that *TYR* was significantly less expressed in the WHITE group compared to NON-DILUTED. This was expected, since *MITF-M* gene, which regulates *TYR* expression, was found to be less expressed in white llamas compared to colored ones (Anello et al., 2019). Our results are also consistent with the observations of Paterson et al. (2015), who demonstrated that partial depletion of *TYR* inhibited melanosome maturation and the expression of genes that regulates melanogenesis, altering mouse coat color.

The expression level of *TYR* was also studied by Munyard (2011), who observed that *TYR*, together with seven other color-related genes, was expressed in a common pattern in alpacas: high

[6]http://grch37.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Table?db=core;g=ENSG00000077498;r=11:88910620-89028927

in black, moderate in bay, and low in white. Similarly, Chen et al. (2012) analyzed the expression of *TYR* in the skin of Jining gray goats and it resulted higher in the dark-gray goats compared to the light-gray ones. Kim et al. (2014) also observed that *TYR* was more expressed in the dark vs. light muzzle of native Korean cows. In our study, the expression in the DILUTED group was in between the other two groups, although differences were not statistically significant. This could be due to a small difference in expression that would need a larger sample to be detectable. Additionally, differences in the degree of dilution of the animals that conformed this group might have influenced this result.

It is commonly known that UTR regions have important regulatory elements that control gene expression (Matoulkova et al., 2012). Therefore, the most important regulatory regions of *TYR* gene were analyzed in this study: the proximal promoter and the LCR. However, all the regulatory elements previously described within these regions were found to be conserved in llamas; the detected variation was located outside the elements. None of the polymorphisms observed in these regions seemed to explain the expression differences observed between white and colored animals.

Unexpectedly c.1-26C > T presented a significantly higher frequency of the C/T genotype in diluted animals compared to the other phenotypic groups. Nevertheless, this polymorphism alone is not causative for the color dilution, since C/T genotype was also observed in non-diluted llamas, where tyrosinase activity is expected to be normal. One possible explanation for this result could be that the same *TYR* mutation under different genetic backgrounds produces different phenotypes, as it was proposed for rabbits of different strains (Aigner et al., 2000). Additionally, different polymorphisms could be contributing together with c.1-26C > T to the final color phenotype. Finally, we cannot exclude that another polymorphism linked to c.1-26C > T, located in regions non-contemplated in this study (like intronic or other regulatory regions) is the actual causal of melanin dilution.

In this study, we have characterized the structure of *TYR* gene and its variation, contributing to the genetic knowledge of the llama. Moreover, we have analyzed the role of *TYR* variation and its expression in different color phenotypes, bringing new information to the understanding of the llama pigmentation mechanisms.

AUTHOR CONTRIBUTIONS

MA, LV-R, and FDR conceived and designed the research. MA and EF performed the experiments. MA, EF, and MD analyzed the data. MA and FDR wrote the manuscript. EF, MD, and LV-R revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by grant PIP-00370 from the National Scientific and Technical Research Council (CONICET) and funds from the Commission of Scientific Research of the Province of Buenos Aires (CIC). FDR is a researcher from CIC.

ACKNOWLEDGMENTS

We thank the llama owners who allowed us to take blood and skin samples, Mr. Carlos Rusconi and Dr. Miragaya and his working team.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00568/full#supplementary-material

REFERENCES

Aigner, B., Besenfelder, U., Müller, M., and Brem, G. (2000). Tyrosinase gene variants in different rabbit strains. *Mamm. Genome* 11, 700–702. doi: 10.1007/s003350010120

Anello, M., Daverio, M. S., Silbestro, M. B., Vidal-Rioja, L., and Di Rocco, F. (2019). Characterization and expression analysis of *KIT* and *MITF-M* genes in llamas and their relation to white coat color. *Anim. Genet.* 50, 1–7. doi: 10.1111/age.12769

Anistoroaei, R., Fredholm, M., Christensen, K., and Leeb, T. (2008). Albinism in the American mink (*Neovison vison*) is associated with a tyrosinase nonsense mutation. *Anim. Genet.* 39, 645–648. doi: 10.1111/j.1365-2052.2008.01788.x

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., et al. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi: 10.1093/nar/gkw1099

Beermann, F., Orlow, S. J., and Lamoreux, M. L. (2004). The Tyr (albino) locus of the laboratory mouse. *Mamm. Genome* 15, 749–758. doi: 10.1007/s00335-004-4002-4008

Belguith-Maalej, S., Kallel, R., Mnif, M., Abid, M., Ayadi, H., and Kacem, H. H. (2013). Association of intronic repetition of SLC26A4 gene with Hashimoto thyroiditis disease. *Genet. Res.* 95, 38–44. doi: 10.1017/S0016672313000037

Benkel, B. F., Rouvinen-Watt, K., Farid, H., and Anistoroaei, R. (2009). Molecular characterization of the Himalayan mink. *Mamm. Genome* 20, 256–259. doi: 10.1007/s00335-009-9177-9176

Blaszczyk, W. M., Arning, L., Hoffmann, K.-P., and Epplen, J. T. (2005). A tyrosinase missense mutation causes albinism in the Wistar rat. *Pigment Cell Res.* 18, 144–145. doi: 10.1111/j.1600-0749.2005.00227.x

Challa, A. K., Boitet, E. R., Turner, A. N., Johnson, L. W., Kennedy, D., Downs, E. R., et al. (2016). Novel hypomorphic alleles of the mouse tyrosinase gene induced by CRISPR-Cas9 nucleases cause non-albino pigmentation phenotypes. *PLoS One* 11:e0155812. doi: 10.1371/journal.pone.0155812

Chandramohan, B., Renieri, C., La Manna, V., and La Terza, A. (2013). The alpaca agouti gene: genomic locus, transcripts and causative mutations of eumelanic and pheomelanic coat color. *Gene* 521, 303–310. doi: 10.1016/j.gene.2013.03.060

Chandramohan, B., Renieri, C., La Manna, V., and La Terza, A. (2015). The alpaca melanocortin 1 receptor: gene mutations, transcripts, and relative levels of expression in ventral skin biopsies. *ScientificWorldJournal* 2015:265751. doi: 10.1155/2015/265751

Chen, W., Wang, H., Dong, B., Dong, Z., Zhou, F., Fu, Y., et al. (2012). Molecular cloning and expression analysis of tyrosinase gene in the skin of Jining gray goat (*Capra hircus*). *Mol. Cell. Biochem.* 366, 11–20. doi: 10.1007/s11010-012-1275-1271

Cieslak, M., Reissmann, M., Hofreiter, M., and Ludwig, A. (2011). Colours of domestication. *Biol. Rev.* 86, 885–899. doi: 10.1111/j.1469-185X.2011.00177.x

Cirera, S., Markakis, M. N., Kristiansen, T., Vissenberg, K., Fredholm, M., Christensen, K., et al. (2016). A large insertion in intron 2 of the TYRP1 gene associated with American Palomino phenotype in American mink. *Mamm. Genome* 27, 135–143. doi: 10.1007/s00335-016-9620-9624

Cransberg, R., and Munyard, K. A. (2011). Polymorphisms detected in the tyrosinase and matp (slc45a2) genes did not explain coat colour dilution in a sample of Alpaca (*Vicugna pacos*). *Small Rumin. Res.* 95, 92–96. doi: 10.1016/j.smallrumres.2010.10.004

Damé, M. C. F., Xavier, G. M., Oliveira-Filho, J. P., Borges, A. S., Oliveira, H. N., Riet-Correa, F., et al. (2012). A nonsense mutation in the tyrosinase gene causes albinism in water buffalo. *BMC Genet.* 13:62. doi: 10.1186/1471-2156-13-62

Daverio, M. S., Rigalt, F., Romero, S., Vidal-Rioja, L., and Di Rocco, F. (2016). Polymorphisms in MC1R and ASIP genes and their association with coat color phenotypes in llamas (*Lama glama*). *Small Rumin. Res.* 144, 83–89. doi: 10.1016/j.smallrumres.2016.08.003

Dobson, L., Reményi, I., and Tusnády, G. E. (2015). CCTOP: a consensus constrained TOPology prediction web server. *Nucleic Acids Res.* 43, W408–W412. doi: 10.1093/nar/gkv451

Feeley, N. L., Bottomley, S., and Munyaerd, K. A. (2011). Three novel mutations in ASIP associated with black fibre in alpacas (Vicugna pacos). *J. Agric. Sci.* 149, 529–538. doi: 10.1017/S0021859610001231

Feeley, N. L., and Munyard, K. A. (2009). Characterisation of the melanocortin-1 receptor gene in alpaca and identification of possible markers associated with phenotypic variations in colour. *Anim. Prod. Sci.* 49, 675–681. doi: 10.1071/AN09005

Frank, E. N., Hick, M. V. H., Gauna, C. D., Lamas, H. E., Renieri, C., and Antonini, M. (2006). Phenotypic and genetic description of fibre traits in South American domestic camelids (*llamas and alpacas*). *Small Rumin. Res.* 61, 113–129. doi: 10.1016/j.smallrumres.2005.07.003

García-Borrón, J. C., and Solano, F. (2002). Molecular anatomy of tyrosinase and its related proteins: beyond the histidine-bound metal catalytic center. *Pigment Cell Res.* 15, 162–173. doi: 10.1034/j.1600-0749.2002.02012.x

Giraldo, P., Martínez, A., Regales, L., Lavado, A., García-Díaz, A., Alonso, A., et al. (2003). Functional dissection of the mouse tyrosinase locus control region identifies a new putative boundary activity. *Nucleic Acids Res.* 31, 6290–6305. doi: 10.1093/nar/gkg793

Guridi, M., Soret, B., Alfonso, L., and Arana, A. (2011). Single nucleotide polymorphisms in the Melanocortin 1 receptor gene are linked with lightness of fibre colour in peruvian alpaca (*Vicugna pacos*). *Anim. Genet.* 42, 679–682. doi: 10.1111/j.1365-2052.2011.02205.x

Imes, D. L., Geary, L. A., Grahn, R. A., and Lyons, L. A. (2006). Albinism in the domestic cat (*Felis catus*) is associated with a tyrosinase (TYR) mutation. *Anim. Genet.* 37, 175–178. doi: 10.1111/j.1365-2052.2005.01409.x

Kanteev, M., Goldfeder, M., and Fishman, A. (2015). Structure-function correlations in tyrosinases. *Protein Sci.* 24, 1360–1369. doi: 10.1002/pro.2734

Kim, S. H., Hwang, S. Y., and Yoon, J. T. (2014). Microarray-based analysis of the differential expression of melanin synthesis genes in dark and light-muzzle Korean cattle. *PLoS One* 9:e96453. doi: 10.1371/journal.pone.0096453

Kwon, B. S., Halaban, R., and Chintamaneni, C. (1989). Molecular basis of mouse Himalayan mutation. *Biochem. Biophys. Res. Commun.* 161, 252–260. doi: 10.1016/0006-291x(89)91588-x

Lamoreux, M. L., Wakamatsu, K., and Ito, S. (2001). Interaction of major coat color gene functions in mice as studied by chemical analysis of eumelanin and pheomelanin. *Pigment Cell Res.* 14, 23–31. doi: 10.1034/j.1600-0749.2001.140105.x

Li, Y. C., Korol, A. B., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21, 991–1007. doi: 10.1093/molbev/msh073

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2-\Delta\Delta CT$ Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lyons, L. A., Imes, D. L., Rah, H. C., and Grahn, R. A. (2005). Tyrosinase mutations associated with siamese and burmese patterns in the domestic cat (*Felis catus*). *Anim. Genet.* 36, 119–126. doi: 10.1111/j.1365-2052.2005.01253.x

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi: 10.1093/nar/gkw1129

Matoulkova, E., Michalova, E., Vojtesek, B., and Hrstka, R. (2012). The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol.* 9, 563–576. doi: 10.4161/rna.20231

Munyard, K. (2011). *Inheritance of White Colour in Alpacas*. Canberra: Rural Industries Research and Development Corporation.

Murisier, F., and Beermann, F. (2006). Genetics of pigment cells: lessons from the tyrosinase gene family. *Histol. Histopathol.* 21, 567–578. doi: 10.14670/HH-21.567

Oetting, W. S. (2000). The tyrosinase gene and oculocutaneous albinism type 1 (OCA1): a model for understanding the molecular biology of melanin formation. *Pigment Cell Res.* 13, 320–325. doi: 10.1034/j.1600-0749.2000.130503.x

Orlow, S. J., Lamoreux, M. L., Pifko-Hirst, S., and Zhou, B. K. (1993). Pathogenesis of the Platinum (cp) mutation, a model for oculocutaneous albinism. *J. Invest. Dermatol.* 101, 137–140. doi: 10.1111/1523-1747.ep12363621

Paterson, E. K., Fielder, T. J., MacGregor, G. R., Ito, S., Wakamatsu, K., Gillen, D. L., et al. (2015). Tyrosinase depletion prevents the maturation of melanosomes in the mouse hair follicle. *PLoS One* 10:e0143702. doi: 10.1371/journal.pone.0143702

Powell, A. J., Moss, M. J., Tree, L. T., Roeder, B. L., Carleton, C. L., Campbell, E., et al. (2008). Characterization of the effect of melanocortin 1 receptor, a member of the hair color genetic locus, in alpaca (*Lama pacos*) fleece color differentiation. *Small Rumin. Res.* 79, 183–187. doi: 10.1016/j.smallrumres.2008.07.025

Reinisalo, M., Putula, J., Mannermaa, E., Urtti, A., and Honkakoski, P. (2012). Regulation of the human tyrosinase gene in retinal pigment epithelium cells: the significance of transcription factor orthodenticle homeobox 2 and its polymorphic binding site. *Mol. Vis.* 18, 38–54. doi: 10.1080/0968776990070202

Sambrook, J., and Russell, D. W. (2001). *Molecular Cloning? a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Saravanaperumal, S. A., Pediconi, D., Renieri, C., and La Terza, A. (2014). Alternative splicing of the sheep MITF gene: novel transcripts detectable in skin. *Gene* 552, 165–175. doi: 10.1016/j.gene.2014.09.031

Schmutz, S. M., Berryere, T. G., Ciobanu, D. C., Mileham, A. J., Schmidtz, B. H., and Fredholm, M. (2004). A form of albinism in cattle is caused by a tyrosinase frameshift mutation. *Mamm. Genome* 15, 62–67. doi: 10.1007/s00335-002-2249-2245

Sjakste, T., Paramonova, N., and Sjakste, N. (2013). Functional significance of microsatellite markers. *Medicina* 49, 505–509. doi: 10.1177/0090591700028006008

SPSS Inc. (2009). *PASW Statistics for Windows, Version 18.0*. Chicago, IL: SPSS Inc.

Utzeri, V. J., Bertolini, F., Ribani, A., Schiavo, G., Dall'Olio, S., and Fontanesi, L. (2016). The albinism of the feral Asinara white donkeys (Equus asinus) is determined by a missense mutation in a highly conserved position of the tyrosinase (TYR) gene deduced protein. *Anim. Genet.* 47, 120–124. doi: 10.1111/age.12386

Vieira, M. L. C., Santini, L., Diniz, A. L., Munhoz, C., and de, F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39, 312–328. doi: 10.1590/1678-4685-GMB-2016-2027

Yokoyama, T., Silversides, D. W., Waymire, K. G., Kwon, B. S., Takeuchi, T., and Overbeek, P. A. (1990). Conserved cysteine to serine mutation in tyrosinase is responsible for the classical albino mutation in laboratory mice. *Nucleic Acids Res.* 18, 7293–7298. doi: 10.1093/nar/18.24.7293

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Anello, Fernández, Daverio, Vidal-Rioja and Di Rocco. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 19 June 2019
doi: 10.3389/fgene.2019.00588

Cellular and Molecular Adaptation of Arabian Camel to Heat Stress

*Abdullah Hoter[1,2], Sandra Rizk[3] and Hassan Y. Naim[2]**

[1] Department of Biochemistry and Chemistry of Nutrition, Faculty of Veterinary Medicine, Cairo University, Giza, Egypt, [2] Department of Physiological Chemistry, University of Veterinary Medicine Hannover, Hanover, Germany, [3] School of Arts and Sciences, Lebanese American University, Beirut, Lebanon

To cope with the extreme heat stress and drought of the desert, the Arabian camel (*Camelus dromedarius*) has developed exceptional physiological and biochemical particularities. Previous reports focused mainly on the physiological features of Arabian camel and neglected its cellular and molecular characteristics. Heat shock proteins are suggested to play a key role in the protein homeostasis and thermotolerance. Therefore, we aim by this review to elucidate the implication of camel HSPs in its physiological adaptation to heat stress and compare them with HSPs in related mammalian species. Correlation of these molecules to the adaptive mechanisms in camel is of special importance to expand our understanding of the overall camel physiology and homeostasis.

Keywords: Arabian camel, heat shock proteins, heat stress, chaperones, desert, adaptation

OPEN ACCESS

Edited by:
Pamela Burger,
University of Veterinary Medicine
Vienna, Austria

Reviewed by:
Pablo Orozco-terWengel,
Cardiff University, United Kingdom
Ajamaluddin Malik,
King Saud University, Saudi Arabia

****Correspondence:***
Hassan Y. Naim
hassan.naim@tiho-hannover.de

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *12 October 2018*
Accepted: *04 June 2019*
Published: *19 June 2019*

Citation:
Hoter A, Rizk S and Naim HY (2019) Cellular and Molecular Adaptation of Arabian Camel to Heat Stress. Front. Genet. 10:588. doi: 10.3389/fgene.2019.00588

INTRODUCTION

Arabian camel (*Camelus dromedarius*), also known as the one humped camel, is a unique large animal belonging to the Camelidae family. This creature is well adapted to endure extreme levels of heat stress and arid conditions of the desert. Nevertheless, it has been used as a valuable source of milk, meat and wool (Kadim et al., 2008; Faye, 2015). Arabian camels exist mainly in the Middle East and parts of tropical and subtropical regions (Dorman, 1984). Historically, camels were widely used as a principal mean of transport of humans and goods between countries, hence known as the ship of desert (Bornstein, 1990). Recently, additional uses of camel have emerged including tourism, racing events and beauty contests, all emphasizing the fact that camel is a precious multipurpose animal (Faye, 2015). Several attempts have been carried out to understand the mystery of camel adaptation and the incredible capability of camel to withstand dehydration, thermal stress and other harsh environmental conditions (Ouajd and Kamel, 2009; Gebreyohanes and Assen, 2017). A distinctive coordination of anatomical, physiological and behavioral criteria is found to play a role in such efficient adaptation.

At the anatomical level, several adaptations have been identified: the one humped camel is provided with a single hump filled with fat rather than the common belief of being filled with water (Mohammed et al., 2005). The high fat content in camel humps serves as an energy store which is used in periods of food limitation (Chilliard, 1989). Camel nostrils have a muscular

nature which allows camel to fully control its opening and closure, thus avoiding sand inhalation in case of sandstorm events (Gebreyohanes and Assen, 2017). The feet of camel are thick and characterized by leathery pads which spread widely on hitting the ground, consequently preventing the animal from sinking into the warm sand. Camel legs are long compared to other desert animals and during walking each two legs move on one side, rocking side-to-side, therefore giving another reason for being nicknamed the ship of desert. Among the interesting internal anatomical features observed in camels is the unique water sac structure in the stomach serving to store water (Allouch, 2016). Interestingly, the anatomical arrangement or distribution of camel arteries and veins help mitigate the high blood temperature of the body reaching the brain, thus protecting the animal from potential brain damage. This mechanism is referred as "selective brain cooling" (Ouajd and Kamel, 2009).

On the other hand, many physiological and behavioral aspects promote the acclimatization of Arabian camels to the extreme heat of the desert. For instance, there is another supporting mechanism to the previously mentioned selective brain cooling known as "adaptive heterothermy." By this mechanism, camel can fluctuate its body temperature between 34 and 42°C, thus minimizing perspiration and avoiding water losses through evaporation (Ouajd and Kamel, 2009). Additionally, camels usually huddle together in order to cool themselves as their body temperature is often less than the surrounding air (Wilson, 1989). Moreover, in the recumbent position, the camel sternum takes a "plate like" conformation permitting more air circulation (Ouajd and Kamel, 2009). Furthermore, camel kidneys are able to efficiently excrete highly concentrated urine consequently tolerating high salt concentrations (Siebert and Macfarlane, 1971). Other physiological particularities include the capability of dromedary to drink huge amounts of water, reaching up to 200 liters at one time to compensate for fluid loss (Ouajd and Kamel, 2009). Red blood cells (RBCs) of camelids are anucleated with an exotic elliptical shape, to presumably facilitate their flow inside blood vessels in dehydrated animals (Al-Swailem et al., 2007; Warda et al., 2014). Moreover, camel platelets can resist high temperatures of 43–45°C which cause marked structural and functional alterations as compared to human platelets. Even higher thermal stress of 50°C that damages human platelets has slight effects on camel cells and does not critically disrupt their function (Al Ghumlas et al., 2008). Surprisingly, camel RBCs possess distinctive membrane phospholipid composition, resulting in a more fluid membrane, and enabling them to bear high osmotic variations without rupturing even in cases of rapid rehydration (Warda and Zeisig, 2000; Warda et al., 2014). Moreover, antibodies in *C. dromedarius* comprise dimeric heavy chains lacking the light chains, however, they display an extensive antigen-binding repertoire (Hamers-Casterman et al., 1993).

All these interesting facts denote further potential unrevealed mechanisms at the molecular level for camel adaptation to various stresses. Indeed, as partially presented in this section, various biomolecules and elements have been characterized in camelids and aided to increase our knowledge about their exceptional cellular homeostasis. However, in the current review we emphasize on camel HSPs as key players in its adaptation to heat stress.

CLASSIFICATION OF FAMILY CAMELIDAE

To get a clearer picture about the organism under study, an obvious simplified classification is presented to avoid confusion with other species within the Camelidae family. The family of Camelidae comprises two major subfamilies, namely Camelinae (Old World Camelids) and Laminae (New World Camelids). The old world camelids include two domesticated species; the dromedary or one humped camel (*C. dromedarius*) and the two humped camel or bactrian camel (*C. bactrianus*). Both species are referred to as large camelids and distributed into different regions of the world. Arabian camel (*C. dromedarius*) is located mainly in the hot areas of Middle East and Africa whereas *C. bactrianus* inhabit the cold zones of Central Asia and China (Al-Swailem et al., 2007; Kadim et al., 2013). The new world camelids comprise four main species located in South America and are commonly known as small camelids. Yet, two species the llama (*Lama glama*) and the alpaca (*Vicugna pacos*) have been domesticated whereas the other two species, namely the guanaco (*L. guanicoe*) and the vicuna (*V. vicugna*) are wild species (Dorman, 1984; Kadim et al., 2008; Faye, 2015). A schematic classification and map distribution of members of the camelidae family is shown in **Figure 1**.

ADAPTATION TO DESERT CONDITIONS IS INTEGRATED IN THE DROMEDARY GENOME

Consistent with their highlighted physiological and anatomical adaptation to desert conditions, dromedary camels have shown interesting findings at the genomic level. In a pioneering work by Wu et al. (2014), they did high quality genome sequencing of three camel species including; *C. dromedarius*, *C. bactrianus*, and alpaca (*V. pacos*). Data achieved through comparative genomic and transcriptomic analyses of those species indicated numerous features with high potential to desert adaptation. For example, the dromedary showed enhanced energy and fat metabolism, water reservation, salt metabolism, osmoregulation and sodium reabsorption. Moreover, stress related genes such as those involved in DNA damage and repair, apoptosis, protein stabilization and immune responses were found superior in terms of accelerated evolution compared to their homologs in cattle species (Wu et al., 2014). In another interesting study, the whole genome sequencing of Iranian dromedaries revealed genetic variations, including single-nucleotide polymorphisms (SNPs) and indels (insertions and deletions) between the compared species. However, the majority of genes associated with stress response were clearly identified in the species under study suggesting efficient adaptation of the Iranian dromedaries to the desert milieu of Iran. Identifying genetic variations including

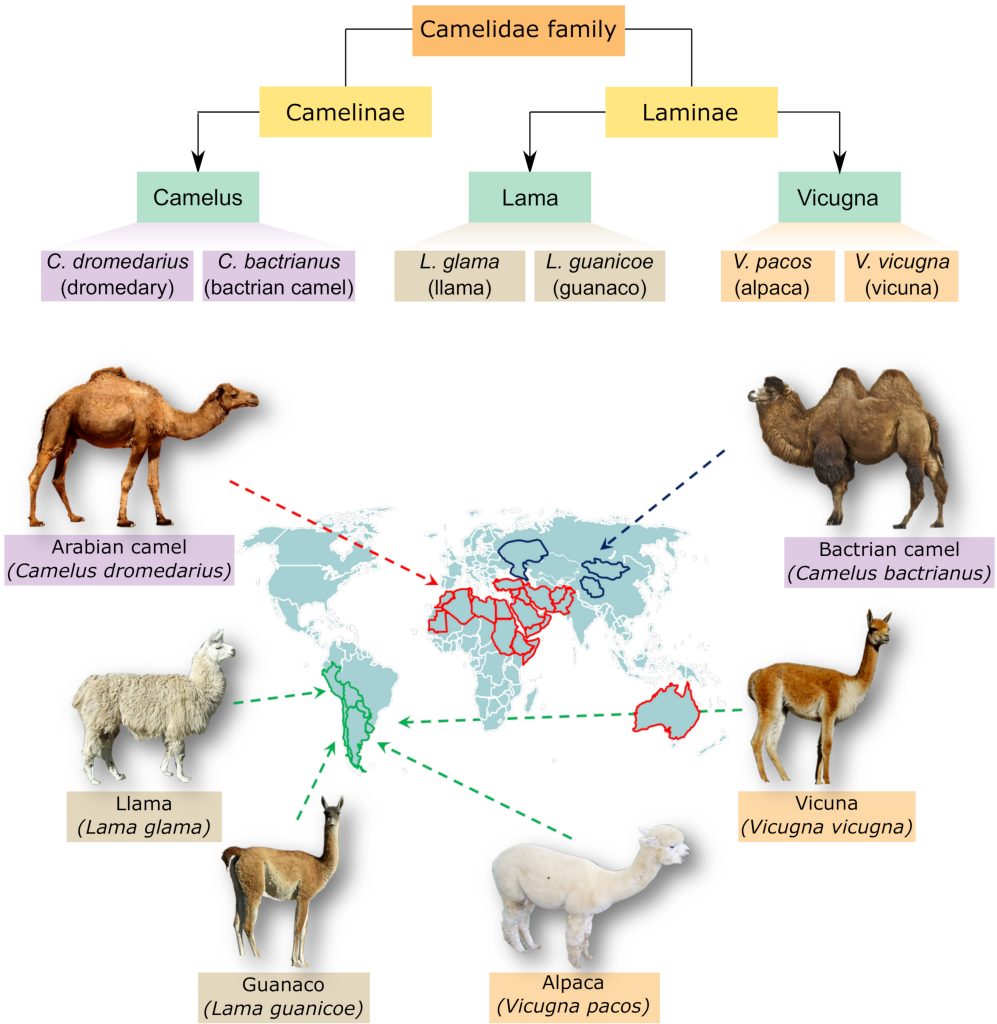


FIGURE 1 | Classification and distribution of camel species. The upper panel demonstrates the genera and species belonging to camelidae family. These include three genera, *Camelus*, *Lama*, and *Vicugna*, which comprise large camel species like Arabian camel (one humped camel) and Bactrian camel (two humped camel) as well as small camelids like llama, alpaca, guanaco and vicuna. *Camelus ferus*, which is a double humped wild Bactrian camel, is not shown for simplicity. The lower panel shows a visual distribution map of different species within camelidae family.

SNPs among native camel species represents a step forward to better understand camel evolution and improves camel breeding programs (Khalkhali-Evrigh et al., 2018).

HEAT SHOCK PROTEINS IN ARABIAN CAMEL IN RELATION TO OTHER ANIMALS

As desert animals, Arabian camels are subjected to extended periods of heat stress which require efficient cellular and molecular buffers. Over the past decade, increasing interest toward camel HSPs has evolved to unravel their potential role in thermotolerance. Most of the studies focused on cloning and characterizing representative HSP members from camel tissues whereas other studies focused on the analysis of HSP expression either in living animals or in mammalian models. Here, we review the highlights of the all respective studies.

HSPA FAMILY (HSP70)

The HSP70 family comprises highly conserved proteins with molecular weight of 70 kDa which are largely known to resist heat as well as several stresses. In humans, the family HSP70 comprises thirteen members which share high sequence and structural homology, however, may vary or overlap in their functions (Daugaard et al., 2007; Kampinga et al., 2009; **Table 1**). The main structural features in HSP70 members include N-terminal domain which has an ATPase activity, a middle domain and a C-terminal domain (**Figure 2A**). The HSP70 family includes housekeeping or continuously expressed proteins in addition to other inducible members. Compared to other HSP families,

TABLE 1 | Various members of HSP70 (Kampinga et al., 2009).

Gene name	Protein name	Alternative name	Human Gene ID
HSPA1A	HSPA1A	HSP70-1; HSP72; HSPA1	3303
HSPA1B	HSPA1B	HSP70-2	3304
HSPA1L	HSPA1L	hum70t; hum70t; Hsp-hom	3305
HSPA2	HSPA2	Heat-shock 70kD protein-2	3306
HSPA5	HSPA5	BiP; GRP78; MIF2	3309
HSPA6	HSPA6	Heat shock 70kD protein 6 (HSP70B')	3310
HSPA7	HSPA7	Heat shock 70kD protein 7	3311
HSPA8	HSPA8	HSC70; HSC71; HSP71; HSP73	3312
HSPA9	HSPA9	GRP75; HSPA9B; MOT; MOT2; PBP74; mot-2	3313
HSPA12A	HSPA12A	FLJ13874; KIAA0417	259217
HSPA12B	HSPA12B	RP23-32L15.1; 2700081N06Rik	116835
HSPA13b	HSPA13b	Stch	6782
HSPA14	HSPA14	HSP70-4; HSP70L1; MGC131990	51182

HSP70 has been mostly studied in correlation to thermal and environmental stresses (Pyza et al., 1997; Ackerman et al., 2000; Kumar et al., 2003).

Early studies on camel lymphocytes revealed high competence of general protein synthesis as compared to humans (Ulmasov et al., 1993). Also, the expression of HSP73 in camel erythrocytes did not show high levels of expression as compared to its expression in lymphocytes. Further molecular analysis of HSPs in camel lymphocytes revealed strong induction of HSP73 upon exposure to thermal stress, whereas the constitutively expressed HSP75 was not equally induced upon the same temperature exposure (Ulmasov et al., 1993).

Further studies on the camel fibroblast cell line model, Dubca, revealed an interesting phenomenon in terms of cell survival. Upon exposure to an elevated temperature of 42°C, the Dubca cells survived the heat stress conditions applied over 48 h and continued to grow while parallel murine fibroblast cells (L929) subjected to the same stress conditions were almost dead upon 24 h exposure (Thayyullathil et al., 2008). Furthermore, the recovery of Dubca cells occurred at a faster rate when compared to the murine cells. Surprisingly, when analyzed by western blot, the HSP70 expression in heat stressed Dubca cells was unexpectedly similar to normal physiological conditions; however, the study revealed the expression of an additional lower size isoform which was suggested to play a role in

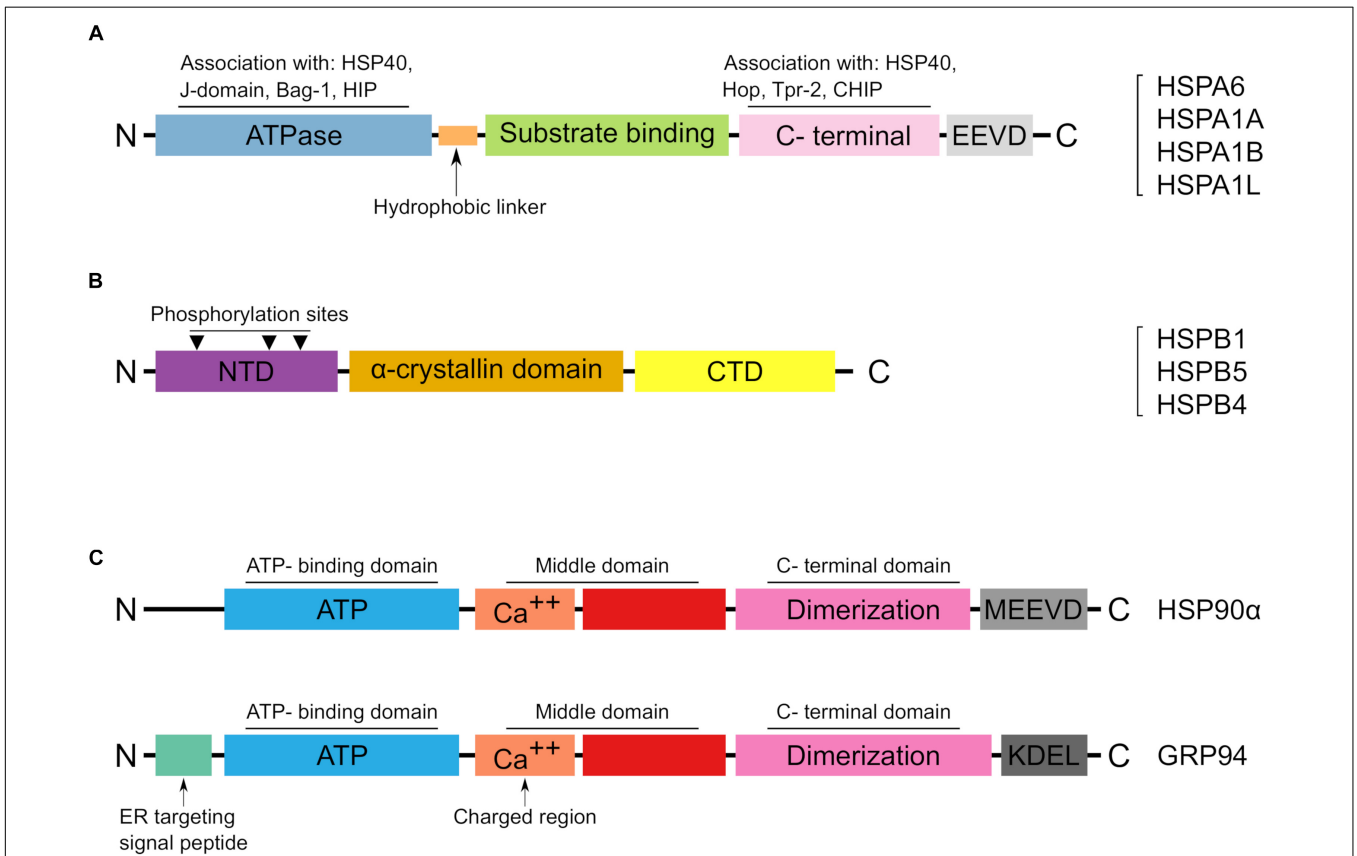


FIGURE 2 | The main structural features of reported HSPs in Arabian camel. **(A)** Schematic representation of the HSP70 structural domains characterized from *Camelus dromedarius*. A list of identified HSP70 members in Arabian camel is shown on the right side. **(B)** The structural characteristics of HSPB proteins in Arabian camel. The reported protein sequences are highly homologous to their human peers and preserve the α-crystalline domain as well as the phosphorylation sites. **(C)** Structural topology of the reported HSP90 members in Arabian camel. The upper panel shows the cytoplasmic HSP90α while the lower panel reveals the ER localized GRP94. HSP90α differs from its ER paralog in lacking both the ER target and retention signals, instead it has a C-terminal MEEVD motif.

the thermotolerance of camel cells (Thayyullathil et al., 2008). Other protein levels of the prosurvival kinase Akt were not altered following 42°C heat stress while the c-Jnk expression was diminished in accordance with the fact that downregulation of c-Jnk has been associated with enhanced thermotolerance (Thayyullathil et al., 2008).

Garbuz et al. (2011) succeeded in mapping a genomic cluster comprising three genes of HSP70 family in *C. dromedarius* namely, *HSPA1A*, *HSPA1L*, and *HSPA1B*. These genes were closely associated with the major histocompatibility complex (MHC) class III region. Two mapped HSP70 genes contained heat shock elements (HSEs) required for heat shock induction, while the third one lacked any (HSEs) sequences. When compared with the corresponding loci in other animals, the camel HSP70 cluster appeared highly conserved and kept similar organization to other species. Notably, the two heat inducible HSPs were tandemly arranged whereas the third constitutive HSP70 gene was located in a reverse orientation. On the other hand, comparison of the regulatory sequences of HSP70 genes in camel and other mammals revealed a greatly conserved arrangement, whereby the sites of transcription factors were localized 250 bp upstream region followed by *NF-Y* and *Sp1* binding sites. Interestingly, the three HSP70 genes were expressed in blood and muscle of camel under both physiological and heat stress conditions indicating a potential thermotolerant role. Overall, the high conservation of HSP70 arrangement in close association with *MHC* locus pinpoints a coordinated functioning of these crucial genes (Garbuz et al., 2011).

Among the HSP70 members, camel HSPA6 attracted a lot of interest. This inducible HSP is strictly regulated, and its expression is reported to be increased after severe cellular stress (Noonan et al., 2007). Interestingly, isolated cDNA of camel HSPA6 showed sequence differences between the cDNA isolated from Saudi Arabia and that isolated from Egypt suggesting the existence of natural variants of HSP70 members among Arabian camel strains (Elrobh et al., 2011; Hoter et al., 2018a). Moreover, in-depth molecular investigation of camel HSPA6 (cHSPA6) revealed that the expression of its isoforms in mammalian cells varies from those exhibited by the human HSPA6 (hHSPA6) homolog (Hoter et al., 2018a). For instance, comparative SDS-PAGE analysis of HSPA6 showed two main isoforms of cHSPA6 while hHSPA6 exhibited 3 closely spaced isoforms at the size level of 70 kDa. The differential mobility pattern on SDS-PAGE suggested differential cellular processing of HSPA6 in the two species. Interestingly, cHSPA6 appeared to contain two-fold increase in *O*-GlcNAc binding sites compared to the human homolog. Further supportive experiments indicated that, unlike the human ortholog, the upper cHSPA6 isoform was comparatively hyperglycosylated and this *O*-glycosylation happens in non-stressed cells as well as hypoxic and heat-stressed cells. The fact that *O*-glycosylation of cytosolic proteins including HSPs promotes stress resistance (Zachara and Hart, 2004) confers additional value to HSPA6 in tolerating heat stress in Arabian camels (Hoter et al., 2018a).

In another interesting study, Sadder et al. (2015) evaluated the HSP expression profiles in Arabian camels under controlled environmental stress. The animals selected were maintained in a constrained climatic chamber and were heat challenged by exposure to 43°C for different time points. HSP expression levels estimated from the collected blood revealed a sharp increase in the mRNA levels of HSP60, HSPA6, HSP105, HSP70, and HSPA1L after 3 h of heat challenge followed by decreased levels after 6 h. Interestingly, the expression of camel HSPs including HSPA6 rebounded after 24 h of heat stress (Sadder et al., 2015). Similarly, the fluctuation of HSP70 expression levels has been reported in cattle where the HSP70 gene expression revealed an initial increase in expression within 1–2 h, followed by a down-regulation after 8 h (Collier et al., 2006; Sadder et al., 2015). In fact, tracking the expression of HSPs under stress conditions in variant animals reflects an outline about the cellular fate in terms of apoptosis/death, or alternatively, stress resistance and cell survival. Also, linking the HSP expression data with animal performance and productivity would help in genetic selection of heat resistant phenotypes (Sadder et al., 2015).

Other recent studies performed on camel cells showed differential tolerance to heat stress (Saadeldin et al., 2018). Both camel oocytes and cumulus granulosa cells were exposed to high temperature of 45°C for 2 or 20 h, representing acute and chronic heat stress conditions. Camel oocytes revealed lower resistance to the applied acute thermal stress unlike cumulus cells which tolerated both acute and chronic stresses. Remarkably, the analysis of mRNA transcripts in both cell types demonstrated significant increases in the expression of *HSP70* and *HSP90* in cumulus compared to their expression in oocytes, indicating that the induction of HSP70 and HSP90 contributes to the preferential enhanced tolerance of cumulus cells to either acute or chronic stress (Saadeldin et al., 2018).

HSPB FAMILY (SMALL HEAT SHOCK PROTEINS, sHSPs)

Small heat shock proteins (sHSPs) are chaperones of small molecular weight, ranging from 12 to 43 kDa. These molecules play a key role in cellular stress resistance and are widely expressed in many cell and tissue types (Bakthisaran et al., 2015). According to their distribution in variant tissues, sHSPs have been classified into two classes: class I and class II (Taylor and Benjamin, 2005). Class I sHSPs includes proteins of ubiquitous expression in almost all tissue types like HSPB1, HSPB5, HSPB6, and HSPB8 while class II sHSPs includes members of target tissue distribution such as HSPB2, HSPB3, CRYAA, HSPB7, HSPB9, and HSPB10 (Taylor and Benjamin, 2005). sHSPs are distinguished from other large molecular weight HSPs by their ATP-independent activity (Basha et al., 2012). Structurally, members of sHSPs share the highly conserved "α-crystallin domain" (ACD) which is considered the hallmark of sHSPs (Basha et al., 2012; **Figure 2B**). A recent classification of human HSPs has designated the name HSPB for sHSP members as shown in **Table 2** (Kampinga et al., 2009).

So far, two members within sHSPs family have been identified in Arabian camel; HSPB5 or (αB-crystallin, CRYAB) and HSPB1

TABLE 2 | Members of HSPB (small heat shock proteins) family (Kampinga et al., 2009).

Gene name	Protein name	Alternative name	Human Gene ID
HSPB1	HSPB1	CMT2F; HMN2B; HSP27; HSP28; HSP25; HS.76067; DKFZp586P1322	3315
HSPB2	HSPB2	MKBP; HSP27; Hs.78846; LOH11CR1K; MGC133245	3316
HSPB3	HSPB3	HSPL27	8988
HSPB4	HSPB4	crystallin alpha A; CRYAA, CRYA1	1409
HSPB5	HSPB5	crystallin alpha B, CRYAB; CRYA2	1410
HSPB6	HSPB6	HSP20; FLJ32389	126393
HSPB7	HSPB7	cvHSP; FLJ32733; DKFZp779D0968	27129
HSPB8	HSPB8	H11; HMN2; CMT2L; DHMN2; E2IG1; HMN2A; HSP22	26353
HSPB9	HSPB9	FLJ27437	94086
HSPB10	HSPB10	ODF1; ODF; RT7; ODF2; ODFP; SODF; ODF27; ODFPG; ODFPGA; ODFPGB; MGC129928; MGC129929	4956
HSPB11	HSPB11	HSP16.2; C1orf41; PP25	51668

or HSP27 (Manee et al., 2017; Hoter et al., 2018a). Molecular investigations of camel CRYAB showed that its coding cDNA contains 528 bp encoding a protein of 175 amino acid residues. Expression analysis of the recombinant cCRYAB by SDS-PAGE reflected a protein band with molecular mass of 25 kDa while confocal microscopic examination of the expressed protein revealed dominant cytoplasmic localization (Hoter et al., 2018a). The cDNA of camel CRYAB and its deduced amino acid sequence showed high similarity and identity with human and other animals. Moreover, camel CRYAB possess the conservative alpha crystalline domain and the putative phosphorylation sites at Ser19, Ser45, and Ser59 (Warda et al., 2014; Hoter et al., 2018a). Phosphorylation of these sites by the MAPK kinase MKK6 potentiates the cytoprotective and chaperone activity of CRYAB in terms of counteracting stress, induced protein aggregation and stabilization of partially denatured or misfolded proteins (Hoover et al., 2000; Bakthisaran et al., 2016).

On the other hand, the camel HSPB1 (HSP27) has a cDNA with open reading frame (ORF) of 606 bp which encodes a protein of 201 amino acids (Manee et al., 2017). The mRNA expression levels of *HSPB1* showed ubiquitous, however, differential expression in various tissues. In non-stressed conditions, the examined camel tissues showed the highest expression of *HSPB1* mRNA in esophagus, skin, and heart compared to the lowest expression in the brain, spleen, and stomach tissues (Manee et al., 2017). When heat stressed for long time at 42°C, the camel skin fibroblast cells (SACAS) showed notable upregulation of HSPB1 following 6 h of incubation compared to control cells incubated at 37°C. These findings indicate that induced expression of sHSPs in Arabian camel is both tissue specific and time dependent (Manee et al., 2017).

Warda et al. (2014) performed an interesting comparative study of camel and rat proteomes in multiple tissues. The comparative proteomic analysis demonstrated marked overexpression of camel CRYAB in camel heart with seven-fold increase as compared to rat heart. This relative increase in cardiac CRYAB expression can be explained considering- the high cytoprotective demand and protein anti-aggregative activity offered by the ATP-independent chaperone. As a consequence, camel heart can utilize minimum energy to withstand the risk of stress induced protein misfolding or aggregation (Warda et al., 2014). Another fruitful outcome from the abundant CRYAB expression in camel heart is promoting structural integrity and providing extra protective roles against dehydration and sudden rehydration in the harsh desert milieu (Warda et al., 2014). Further proteomic analysis of Arabian camel hump revealed excess array of adipose tissue associated cytoskeletal proteins such as actin, tubulin and vimentin in addition to heat shock proteins including HSP27 and HSP70. These cytoskeletal proteins and heat shock proteins provide structural integrity and ensure efficient heat stress tolerance and general protein homeostasis (Warda et al., 2014).

HSPC (HSP90 FAMILY)

HSP90 family is a class of HSPs that has an estimated molecular weight of 90 kDa (Csermely et al., 1998; Chen et al., 2005). This family comprises four major members which are well conserved in higher eukaryotic species; these include HSP90α, HSP90β, GRP94, and TRAP1. HSP90 family has been recently named HSPC according to the guidelines of the HUGO Gene Nomenclature Committee (HGNC) (Kampinga et al., 2009) as presented in **Table 3**, however, they are still eminent with the old name HSP90. Due to their importance and implication in various cellular as well as pathological events (Hoter et al., 2018c), HSP90 isoforms are distributed in crucial cellular compartments. For instance, the two members HSP90α/β are localized in the cytoplasm (Li and Buchner, 2012; Schopf et al., 2017), GRP94 resides in the endoplasmic reticulum (Yang and Li, 2005) and TRAP1 has a mitochondrial preference (Altieri et al., 2012). The vital physiological processes maintained by HSP90 members include protein folding, cell proliferation and differentiation, apoptotic processes, hormone signaling and cell cycle control (Csermely et al., 1998; Hoter et al., 2018c). Additionally, HSP90 candidates are linked to many pathologies such as cancer, inflammation and neurodegenerative diseases (Whitesell and Lindquist, 2005; Luo et al., 2010; Sevin et al., 2015).

TABLE 3 | Different candidates of HSP90 family (Kampinga et al., 2009).

Gene name	Protein name	Alternative name	Human Gene ID
HSPC1	HSPC1	HSP90AA1; HSPN; LAP2; HSP86; HSPC1; HSPCA; HSP89; HSP90;HSP90A; HSP90N; HSPCAL1; HSPCAL4; FLJ31884	3320
HSPC2	HSPC2	HSP90AA2; HSPCA; HSPCAL3; HSP90ALPHA	3324
HSPC3	HSPC3	HSP90AB1; HSPC2; HSPCB; D6S182; HSP90B; FLJ26984; HSP90-BETA	3326
HSPC4	HSPC4	HSP90B1;ECGP; GP96; TRA1; GRP94; endoplasmin	7184
HSPC5	HSPC5	TRAP1; HSP75; HSP90L	10131

In *C. dromedarius*, two candidates of HSP90 have been characterized on a molecular level. The first member that was characterized is HSP90α (Saeed et al., 2015). This cytoplasmic chaperone is induced by variant stresses and is considered as the major form (Sreedhar et al., 2004). Although it displays a high homology at the protein level with its cytoplasmic constitutive homolog, HSP90β, there exists functional differences between the two HSP90 isoforms (Sreedhar et al., 2004; Hoter et al., 2018c). The coding cDNA of camel HSP90α encodes a protein of 733 amino acid residues and shares high similarity and identity with other mammalian HSP90α. Comparative protein sequence analysis reveals over 85% identity between camel and other animals including cattle, horse, dog, cat and human (Saeed et al., 2015). Also, the structural architecture of HSP90 family has been well preserved in the camel HSP90α; these include the N-terminal domain, the middle domain and C-terminal domains with its peptide MEEVD motif. The main structural features of camel HSP90 is demonstrated in **Figure 2C**.

The second HSP90 member characterized in Arabian camel is endoplasmin or glucose regulated protein, GRP94 (Hoter et al., 2018b). This ER resident chaperone is an essential member of the ER quality control machinery. It helps protein folding of nascent polypeptides, binds calcium by its calcium binding domain, interacts with other ER chaperones and targets misfolded proteins to the ER associated degradation pathway (ERAD) (Marzec et al., 2012). Interestingly, camel endoplasmin shares 100% protein identity with that of human and more than 98% with other close mammals (Hoter et al., 2018b). As a consequence, the structural characteristics and posttranslational modifications of the ER chaperone resemble those in human. For instance, camel GRP94 contains a signal sequence at its N-terminal, comprising the first 21 amino acid residues which is cleaved co-translationally to give the mature form of the protein. Also, cGRP94 contains the classical domains: N-terminal domain (NTD), acidic linker domain (LD), middle do- main (MD) and the C- terminal domain (CTD) besides the well-known ER retention motif KDEL (Hoter et al., 2018b). The high protein and structural conservation of HSP90 members among mammals including camel is interesting and indicates the global significance of these HSPs in higher eukaryotes.

CONCLUSION AND FUTURE PERSPECTIVES

For a long time, the Arabian camel has been appreciated as a mean of transport in the desert and as a source of food in terms of meat and milk. Despite the conditions of drought, food limitations and extreme temperatures these interesting creatures can reproduce and function normally without any physiological impairment, a phenomenon that attracts our scientific curiosity to investigate and decipher. Though valuable effort has been done to elucidate the physiological and biochemical secrets of Arabian camel, more in-depth investigations of the molecular adaptation mechanisms in these animals are needed. HSPs, as critical elements in stress response and thermotolerance particularly in desert animals, are worthy candidates to study and evaluate. In this regard, we reviewed the current knowledge in the field of camel HSPs and we strongly encourage further molecular studies of other undeciphered members. Molecular studies of camel HSPs would help expand our knowledge about HSPs and consolidate the interesting physiological phenomena in Arabian camel.

AUTHOR CONTRIBUTIONS

AH and HN conceived the review topic. AH wrote the first draft and designed the figures. SR and HN edited and approved the final version of the review.

FUNDING

Work cited in this review from the Naim laboratory was funded by intramural funds of the University of Veterinary Medicine Hannover, Hanover, Germany. AH was supported by a scholarship from the German Academic Exchange Service (DAAD), Bonn, Germany. This publication was supported by the Deutsche Forschungsgemeinschaft and the University of Veterinary Medicine Hannover, Foundation within the funding programme Open Access Publishing.

REFERENCES

Ackerman, P. A., Forsyth, R. B., Mazur, C. F., and Iwama, G. K. (2000). Stress hormones and the cellular stress response in salmonids. *Fish Physiol. Biochem.* 23, 327–336.

Al Ghumlas, A. K., Gader, A. A., Hussein, M., AlHaidary, A. A., and White, J. G. (2008). Effects of heat on camel platelet structure and function - a comparative study with humans. *Platelets* 19, 163–171. doi: 10.1080/09537100701882061

Allouch, G. (2016). Anatomical study of the water cells area in the dromedary camels rumen (*Camelus dromedarius*). *Nova J. Med. Biol. Sci.* 5, 1–4. doi: 10.20286/nova-jmbs-050183

Al-Swailem, A. M., Al-Busadah, K. A., Shehata, M. M., Al-Anazi, I. O., and Askari, E. (2007). Classification of Saudi Arabian camel (*Camelus dromedarius*) subtypes based on RAPD technique. *J. Food Agric. Environ.* 5, 143–148.

Altieri, D. C., Stein, G. S., Lian, J. B., and Languino, L. R. (2012). TRAP-1, the mitochondrial Hsp90. *Biochim. Biophys. Acta* 1823, 767–773. doi: 10.1016/j.bbamcr.2011.08.007

Bakthisaran, R., Akula, K. K., Tangirala, R., and Rao, ChM. (2016). Phosphorylation of αB-crystallin: role in stress, aging and patho-physiological conditions. *Biochim. Biophys. Acta* 1860, 167–182. doi: 10.1016/j.bbagen.2015.09.017

Bakthisaran, R., Tangirala, R., and Rao, C. M. (2015). Small heat shock proteins: role in cellular functions and pathology. *Biochim. Biophys. Acta* 1854, 291–319. doi: 10.1016/j.bbapap.2014.12.019

Basha, E., O'Neill, H., and Vierling, E. (2012). Small heat shock proteins and α-crystallins: dynamic proteins with flexible functions. *Trends Biochem. Sci.* 37, 106–117. doi: 10.1016/j.tibs.2011.11.005

Bornstein, S. (1990). The ship of the desert. The dromedary camel (*Camelus dromedarius*), a domesti- cated animal species well adapted to extreme conditions of aridness and heat camelus spp origins of the camelidae. *Rangifer* 10, 231–236. doi: 10.7557/2.10.3.860

Chen, B., Piel, W. H., Gui, L., Bruford, E., and Monteiro, A. (2005). The HSP90 family of genes in the human genome: insights into their divergence and evolution. *Genomics* 86, 627–637. doi: 10.1016/j.ygeno.2005.08.012

Chilliard, Y. (1989). "Particularités du métabolisme des lipides et du métabolisme énergétique chez le dromadaire," in *Séminaire sur la digestion, la nutrition et l'alimentation du dromadaire Options Méditerranéennes: Série A. Séminaires Méditerranéens*, ed. Tisserand J.-L. (Zaragoza: CIHEAM), 101–110

Collier, R. J., Stiening, C. M., Pollard, B. C., VanBaale, M. J., Baumgard, L. H., Gentry, P. C., et al. (2006). Use of gene expression microarrays for evaluating environmental stress tolerance at the cellular level in cattle. *J. Anim. Sci.* 84, (Suppl_13) E1–E13

Csermely, P., Schnaider, T., Soti, C., Prohászka, Z., and Nardai, G. (1998). The 90-kDa molecular chaperone family: structure, function, and clinical applications. a comprehensive review. *Pharmacol. Ther.* 79, 129–168. doi: 10.1016/S0163-7258(98)00013-8

Daugaard, M., Rohde, M., and Jäättelä, M. (2007). The heat shock protein 70 family: highly homologous proteins with overlapping and distinct functions. *FEBS Lett.* 581, 3702–3710. doi: 10.1016/j.febslet.2007.05.039

Dorman, A. E. (1984). Aspects of the husbandry and management of the genus Camelus. *Br. Vet. J.* 140, 616–633. doi: 10.1016/0007-1935(84)90013-7

Elrobh, M. S., Alanazi, M. S., Khan, W., Abduljaleel, Z., Al-Amri, A., and Bazzi, M. D. (2011). Molecular cloning and characterization of cDNA encoding a putative stress-induced heat-shock protein from *Camelus dromedarius*. *Int. J. Mol. Sci.* 12, 4214–4236. doi: 10.3390/ijms12074214

Faye, B. (2015). Role, distribution and perspective of camel breeding in the third millennium economies. *Emirates J. Food Agric.* 27, 318–327. doi: 10.9755/ejfa.v27i4.19906

Garbuz, D. G., Astakhova, L. N., Zatsepina, O. G., Arkhipova, I. R., Nudler, E., and Evgen'ev, M. B. (2011). Functional organization of hsp70 cluster in camel (*Camelus dromedarius*) and other mammals. *PLoS One* 6:e27205. doi: 10.1371/journal.pone.0027205

Gebreyohanes, M. G., and Assen, A. M. (2017). Adaptation mechanisms of camels (*Camelus dromedarius*) for desert environment: a review. *J. Vet. Sci. Technol.* 8, 6–10. doi: 10.4172/2157-7579.1000486

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, C., Songa, E. B., et al. (1993). Naturally occurring antibodies devoid of light chains. *Nature* 363, 446–448. doi: 10.1038/363446a0

Hoover, H. E., Thuerauf, D. J., Martindale, J. J., and Glembotski, C. C. (2000). alpha B-crystallin gene induction and phosphorylation by MKK6-activated p38. A potential role for alpha B-crystallin as a target of the p38 branch of the cardiac stress response. *J. Biol. Chem.* 275, 23825–23833. doi: 10.1074/jbc.M003864200

Hoter, A., Amiri, M., Prince, A., Amer, H., Warda, M., and Naim, H. Y. (2018a). Differential glycosylation and modulation of camel and human HSP isoforms in response to thermal and hypoxic stresses. *Int. J. Mol. Sci.* 19:402. doi: 10.3390/ijms19020402

Hoter, A., Amiri, M., Warda, M., and Naim, H. Y. (2018b). Molecular cloning, cellular expression and characterization of Arabian camel (*Camelus dromedarius*) endoplasmin. *Int. J. Biol. Macromol.* 117, 574–585. doi: 10.1016/j.ijbiomac.2018.05.196

Hoter, A., El-Sabban, M. E., and Naim, H. Y. (2018c). The HSP90 family: structure, regulation, function, and implications in health and disease. *Int. J. Mol. Sci.* 19:E2560. doi: 10.3390/ijms19092560

Kadim, I. T., Mahgoub, O., Al-marzooqi, W., Khalaf, S. K., and Raiymbek, G. (2013). Composition, quality and health aspects of the dromedary (*Camelus dromedarius*) and bactrian (Camelus bacterianus) camel meats: a review. *J. Agric. Mar. Sci.* 18, 7–24.

Kadim, I. T., Mahgoub, O., and Purchas, R. W. (2008). A review of the growth, and of the carcass and meat quality characteristics of the one-humped camel (Camelus dromedaries). *Meat Sci.* 80, 555–569. doi: 10.1016/j.meatsci.2008.02.010

Kampinga, H. H., Hageman, J., Vos, M. J., Kubota, H., Tanguay, R. M., and Bruford, E. A, et al. (2009). Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress Chaperones* 14, 105–111. doi: 10.1007/s12192-008-0068-7

Khalkhali-Evrigh, R., Hafezian, S. H., Hedayat-Evrigh, N., Farhadi, A., and Bakhtiarizadeh, M. R. (2018). Genetic variants analysis of three dromedary camels using whole genome sequencing data. *PLoS One* 13:e0204028. doi: 10.1371/journal.pone.0204028

Kumar, Y., Chawla, A., and Tatu, U. (2003). Heat shock protein 70 as a biomarker of heat stress in a simulated hot cockpit. *Aviat. Space Environ. Med.* 74, 711–716.

Li, J., and Buchner, J. (2012). Structure, function and regulation of the hsp90 machinery. *Biomed. J.* 36, 106–117. doi: 10.4103/2319-4170.113230

Luo, W., Sun, W., Taldone, T., Rodina, A., and Chiosis, G. (2010). Heat shock protein 90 in neurodegenerative diseases. *Mol. Neurodegener.* 5, 1–8. doi: 10.1186/1750-1326-5-24

Manee, M. M., Alharbi, S. N., Algarni, A. T., Alghamdi, W. M., Altammami, M. A., Alkhrayef, M. N., et al. (2017). Molecular cloning, bioinformatics analysis, and expression of small heat shock protein beta-1 from *Camelus dromedarius*, Arabian camel. *PLoS One* 12:e0189905. doi: 10.1371/journal.pone.0189905

Marzec, M., Eletto, D., and Argon, Y. (2012). GRP94: an HSP90-like protein specialized for protein folding and quality control in the endoplasmic reticulum. *Biochim. Biophys. Acta* 1823, 774–787. doi: 10.1016/j.bbamcr.2011.10.013

Mohammed, B., Faulconnier, Y., Tabarani, A., Sghiri, A., Faye, B., and Chilliard, A (2005). Effects of feeding level on body weight, hump size, lipid content and adipocyte volume in the dromedary camel. *Anim. Res.* 54, 383–393. doi: 10.1051/animres:2005029

Noonan, E. J., Place, R. F., Giardina, C., and Hightower, L. E. (2007). Hsp70B' regulation and function. *Cell Stress Chaperones* 12, 393–402. doi: 10.1379/CSC-278e.1

Ouajd, S., and Kamel, B. (2009). Physiological particularities of dromedary (*Camelus dromedarius*) and experimental implications. *Scand. J. Lab. Anim. Sci.* 36, 19–29. doi: 10.23675/sjlas.v36i1.165

Pyza, E., Mak, P., Kramarz, P., and Laskowski, R. (1997). Heat shock proteins (HSP70) as biomarkers in ecotoxicological studies. *Ecotoxicol. Environ. Saf.* 38, 244–251. doi: 10.1006/eesa.1997.1595

Saadeldin, I. M., Swelum, A. A.-A., Elsafadi, M., Mahmood, A., Alfayez, M., and Alowaimer, A. N. (2018). Differences between the tolerance of camel oocytes and cumulus cells to acute and chronic hyperthermia. *J. Therm. Biol.* 74, 47–54. doi: 10.1016/j.jtherbio.2018.03.014

Sadder, M. T., Migdadi, H. M., Zakri, A. M., Abdoun, K. A., Samara, E. M., Okab, A. B., et al. (2015). Expression analysis of heat shock proteins in dromedary camel (*Camelus dromedarius*). *J. Camel Pract. Res.* 22, 19–24. doi: 10.5958/2277-8934.2015.00003.X.

Saeed, H., Shalaby, M., Embaby, A., Ismael, M., Pathan, A., Ataya, F., et al. (2015). The arabian camel *Camelus dromedarius* heat shock protein 90α: cDNA cloning, characterization and expression. *Int. J. Biol. Macromol.* 81, 195–204. doi: 10.1016/j.ijbiomac.2015.07.058

Schopf, F. H., Biebl, M. M., and Buchner, J. (2017). The HSP90 chaperone machinery. *Nat. Rev. Mol. Cell Biol.* 18, 345–360. doi: 10.1038/nrm.2017.20

Sevin, M., Girodon, F., Garrido, C., and De Thonel, A. (2015). HSP90 and HSP70: implication in inflammation processes and therapeutic approaches for myeloproliferative neoplasms. *Mediators Inflamm.* 2015:970242. doi: 10.1155/2015/970242

Siebert, B. D., and Macfarlane, W. V. (1971). Water turnover and renal function of dromedaries in the desert. *Physiol. Zool.* 44, 225–240. doi: 10.1086/physzool.44.4.30152494

Sreedhar, A. S., Kalmár, É., Csermely, P., and Shen, Y. F. (2004). Hsp90 isoforms: functions, expression and clinical importance. *FEBS Lett.* 562, 11–15. doi: 10.1016/S0014-5793(04)00229-7

Taylor, R. P., and Benjamin, I. J. (2005). Small heat shock proteins: a new classification scheme in mammals. *J. Mol. Cell. Cardiol.* 38, 433–444. doi: 10.1016/j.yjmcc.2004.12.014

Thayyullathil, F., Chathoth, S., Hago, A., Wernery, U., Patel, M., and Galadari, S. (2008). Investigation of heat stress response in the camel fibroblast cell line dubca. *Ann. N. Y. Acad. Sci.* 1138, 376–384. doi: 10.1196/annals.1414.039

Ulmasov, H. A., Karaev, K. K., Lyashko, V. N., and Evgen'ev, M. B. (1993). Heat-shock response in camel (*Camelus dromedarius*) blood cells and adaptation to hyperthermia. *Comp. Biochem. Physiol. B.* 106, 867–872 doi: 10.1016/0305-0491(93)90043-5

Warda, M., Prince, A., Kim, H. K., Khafaga, N., Scholkamy, T., Linhardt, R. J., et al. (2014). Proteomics of old world camelid (*Camelus dromedarius*): better understanding the interplay between homeostasis and desert environment. *J. Adv. Res.* 5, 219–242. doi: 10.1016/j.jare.2013.03.004

Warda, M., and Zeisig, R. (2000). Phospholipid- and fatty acid-composition in the erythrocyte membrane of the one-humped camel [*Camelus dromedarius*] and its influence on vesicle properties prepared from these lipids. *Dtsch. Tierarztl. Wochenschr.* 107, 368–373.

Whitesell, L., and Lindquist, S. L. (2005). HSP90 and the chaperoning of cancer. *Nat. Rev. Cancer* 5, 761–772. doi: 10.1038/nrc1716

Wilson, R. T. (1989). *Ecophysiology of the Camelidae and Desert Ruminants*. New York, NY: Springer-Verlag.

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun*. 5:5188. doi: 10.1038/ncomms6188

Yang, Y., and Li, Z. (2005). Roles of heat shock protein gp96 in the ER quality control: redundant or unique function? *Mol. Cells* 20, 173–182. doi: 10.1016/j.molcel.2005.10.002

Zachara, N. E., and Hart, G. W. (2004). O-GlcNAc a sensor of cellular state: the role of nucleocytoplasmic glycosylation in modulating cellular function in response to nutrition and stress. *Biochim. Biophys. Acta* 1673, 13–28. doi: 10.1016/j.bbagen.2004.03.016

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor and reviewer PO-t declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

Copyright © 2019 Hoter, Rizk and Naim. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 June 2019
doi: 10.3389/fgene.2019.00583

Chromosomal Localization of Candidate Genes for Fiber Growth and Color in Alpaca (*Vicugna pacos*)

Mayra N. Mendoza[1], Terje Raudsepp[2], Fahad Alshanbari[2], Gustavo Gutiérrez[1]* and F. Abel Ponce de León[3]*

[1] Programa de Mejoramiento Animal, Universidad Nacional Agraria La Molina, Lima, Peru, [2] Molecular Cytogenetics and Genomics Laboratory, Texas A&M University, College Station, TX, United States, [3] Department of Animal Science, University of Minnesota, Minneapolis, MN, United States

The alpaca (*Vicugna pacos*) is an economically important and cultural signature species in Peru. Thus, molecular genomic information about the genes underlying the traits of interest, such as fiber properties and color, is critical for improved breeding and management schemes. Current knowledge about the alpaca genome, particularly the chromosomal location of such genes of interest is limited and lags far behind other livestock species. The main objective of this work was to localize alpaca candidate genes for fiber growth and color using fluorescence *in situ* hybridization (FISH). We report the mapping of candidate genes for fiber growth *COL1A1*, *CTNNB1*, *DAB2IP*, *KRT15*, *KRTAP13-1*, and *TNFSF12* to chromosomes 16, 17, 4, 16, 1, and 16, respectively. Likewise, we report the mapping of candidate genes for fiber color *ALX3*, *NCOA6*, *SOX9*, *ZIC1*, and *ZIC5* to chromosomes 9, 19, 16, 1, and 14, respectively. In addition, since *KRT15* clusters with five other keratin genes (*KRT31*, *KRT13*, *KRT9*, *KRT14*, and *KRT16*) in scaffold 450 (Vic.Pac 2.0.2), the entire gene cluster was assigned to chromosome 16. Similarly, mapping *NCOA6* to chromosome 19, anchored scaffold 34 with 8 genes, viz., *RALY*, *EIF2S2*, *XPOTP1*, *ASIP*, *AHCY*, *ITCH*, *PIGU*, and *GGT7* to chromosome 19. These results are concordant with known conserved synteny blocks between camelids and humans, cattle and pigs.

Keywords: alpaca, chromosomes, FISH, mapping, fiber, color, genes

OPEN ACCESS

Edited by:
Pamela Burger,
University of Veterinary Medicine Vienna, Austria

Reviewed by:
Mohammed Piro,
Agronomic and Veterinary Institute Hassan II, Morocco
Warren Johnson,
Smithsonian Institution, United States

****Correspondence:***
Mayra N. Mendoza
Mayra.Mendoza.cerna@outlook.com
Gustavo Gutiérrez
gustavogr@lamolina.edu.pe

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *29 October 2018*
Accepted: *04 June 2019*
Published: *19 June 2019*

Citation:
Mendoza MN, Raudsepp T, Alshanbari F, Gutiérrez G and Ponce de León FA (2019) Chromosomal Localization of Candidate Genes for Fiber Growth and Color in Alpaca (Vicugna pacos). Front. Genet. 10:583. doi: 10.3389/fgene.2019.00583

INTRODUCTION

The alpaca (*Vicugna pacos*) is a domesticated South American camelid adapted to the Andean climate conditions. They are economically important in Peru as a fiber production species benefiting the small shareholders living in this geographical region (Quispe et al., 2009). Alpaca fiber is highly valued in the international market because of its softness and resistance (Crispín, 2008). Alpacas carry a cultural value because of their historical importance, millenary tradition, ancestral Peruvian identity and unique characteristics derived from their adaptation to the Andean geography and climate (Yucra, 2017). Alpaca meat is highly valued for its high protein and low cholesterol content (Hack, 2001), and continues serving rural population of Altiplano as an important source of protein (Cruz et al., 2017).

Management systems promoting the improvement of alpaca herd productivity have not yet been adopted widely (Quispe et al., 2009). Actual research is orientated to the application of genetic improvement technologies that would decrease fiber diameter, increase fleece weight, and establish uniform color herds (Morante et al., 2009). Genomic selection using single nucleotide polymorphisms (SNPs)-based genotype-phenotype associations, offers the best option presently available. To apply genomic selection in alpacas, it is necessary to identify and map SNPs throughout the genome and associate them with genes that control economic productive traits. In turn, mapping candidate genes already reported in association to color and fiber characteristics, as well as SNPs, will contribute to understanding the organization of the alpaca genome and genome-wide selection of appropriate markers to develop molecular marker microarrays.

Cytogenetic analysis has demonstrated that all camelids share the same chromosome number (2n = 74) with essentially similar chromosome morphology and banding patterns (Hsu and Benirschke, 1967; Taylor et al., 1968; Bianchi et al., 1986). The first camelid chromosome map was based on Zoo-FISH revealing evolutionarily conserved synteny segments across the dromedary, human, cattle and pig (Balmus et al., 2007). This information was instrumental for starting systematic gene mapping in these species and the first cytogenetics maps for the alpaca genome were developed only recently (Avila et al., 2014a,b, 2015). Because of difficulties to unambiguously identify camelid chromosomes (Di Berardino et al., 2006; Avila et al., 2014b), the 230 cytogenetically mapped markers in alpaca (Avila et al., 2014a) will serve as critical references for FISH-mapping new genes and markers.

The aim of this study was to cytogenetically map 11 alpaca candidate genes for fiber growth and coat color to progress the development of alpaca cytogenetic map and chromosomal anchoring the reference sequence.

MATERIALS AND METHODS

Chromosome Preparations

Alpaca chromosome slides were prepared from peripheral blood lymphocytes of normal alpacas according to standard protocols (Raudsepp and Chowdhary, 2008). We used Concanavalin A (Con A from *Canavalia ensiformis*, 20 μg/ml; Sigma Aldrich) as the mitogen, instead of Pokeweed, because Con A stimulates better proliferation of alpaca blood lymphocytes (Avila et al., 2015).

Gene Selection and Primer Design

Genes for cytogenetic mapping were retrieved from publications. Candidate genes regulating fiber growth characteristics, *COL1A1*, *CTNNB1*, *DAB2IP*, *KRT15*, and *TNFSF12* (Fernandez, 2015), and *KRTAP13-1* (Florez, 2016); candidate genes that regulate the expression of fiber color, *NCOA6-agouti* chimera (Chandramohan et al., 2013); *ZIC1*, *ZIC5*, and *SOX9* which conform the neural crest gene regulatory network (Simoes-Costa and Bronner, 2013), and the *ALX3* transcription factor that regulates melanocyte differentiation in striped rodents (Cuthill et al., 2017). Gene specific sequences were retrieved from VicPac 2.0.2 (GCA_000164845.3) at the NCBI (National Center for Biotechnology Information). Since each of the selected genes are members of gene super-families, sequences that characterized these super-families were identified using the BLASTp[1] and Spling[2] tools and manually removed from each gene FASTA sequence. This way unique sequences for each specific gene were obtained. The gene sequences were masked for repeats in RepeatMasker[3]. Gene-specific PCR primers were designed with Primer3 (Untergasser et al., 2012)[4] and Primer-BLAST[5] software packages. The primers were tested by *in silico* PCR[6] and optimized on alpaca genomic DNA.

Overgo primers were designed manually from 36 to 52 bp size sequence within the PCR amplicon. We designed a 24 bp forward primer from the first nucleotide at the 5′ end position of the selected region. The reverse primer was designed starting at the 3′ end of the selected region, ending with 8 nucleotides overlapping the forward primer. The overlapping section and the single strand sections of the forward and reverse primers, contained 50–60 (±5) % GC (we used GC calculator[7]). PCR and overgo primers for each gene are presented in **Table 1**.

Alpaca CHORI-246 Library Screening and BAC DNA Isolation

BAC clones containing sequences of the selected genes were identified as described by Avila et al. (2014b). Briefly, pools of radioactively labeled [(^{32}P) dATP/dCTP] overgo primers were hybridized to CHORI-246 alpaca BAC library[8] filters. Filters were exposed to autoradiography films and positive BAC clones were identified and picked from the library. BACs corresponding to individual genes were identified by PCR with gene-specific primers. BAC DNA was isolated with the Plasmid Midi Kit (Qiagen) and evaluated for quality by electrophoresis in 1% agarose gels.

Probe Labeling, FISH and Microscopy

BAC DNA labeling, hybridizations and signal detection were carried out according to standard protocols (Raudsepp and Chowdhary, 2008). The DNA of individual BACs was labeled with biotin or digoxigenin using DIG- or Biotin-Nick Translation Mix (Roche Diagnostics) and the manufacturer's protocol. Because the known difficulties to unambiguously identify camelid chromosomes, we consulted Zoo-FISH data (Balmus et al., 2007) and the 230-marker cytogenetic map (Avila et al., 2014a) to infer the most probable chromosome location for each candidate gene.

[1]https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins

[2]https://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi?textpage=online&level=form

[3]http://www.repeatmasker.org/

[4]http://bioinfo.ut.ee/primer3-0.4.0/primer3/

[5]https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome

[6]https://genome.ucsc.edu/cgi-bin/hgPcr

[7]http://www.sciencelauncher.com/oligocalc.html

[8]https://bacpacresources.org/libraries.htm

TABLE 1 | Gene specific PCR and overgo primers.

Gene symbol	Identified BAC clones	PCR primer 5′-3′	PCR product size (bp)	Overgo primer sequence 5′-3′
ALX3	**115I10**	F: TATGTCTCCGTACTCCCACTCTC R: GGAGACTTATAGTCGTCATCTGG	161	F: GCTCTAGGGGGCCACAGCTTTGAG R: CGTCATCTGGGGAGGGCTCAAAGC
COL1A1	198E13 **204B18** 264O11 271L20 295O17	F: CCATTGGTAGTGTTGGTGCT R: AGGGAAGCCTCTTTCTCCTC	365	F: GCCCTGTTGGCAAAGAAGGCAGCA R: TCACCACGAGGACCTTTGCTGCCT
CTNNB1	129B09 **150A21**	F: ATCCCAGCTATCGTTCTTTTCA R: CCTACCAACCCAAGTCTTTCTG	300	F: CACTCCGGTGGATACGGACAGGAT R: GGTCCATACCCAAGGCATCCTGTC
DAB2IP	**101B06**	F: TACTGAGAACGGCGAGTTCA R: AAAGCTCAGCCTCTCTCTCG	107	F: GAACGGCGAGTTCAGAAACAGCAGCAA R: CGTGCCTGGGACACTTGAATTGCTGCT
KRT15	263E22 **268A9** 274A22	F: GGCAAAGTCCGCATCAATGTT R: ATGCCAAGCAGCCAACTAGG	218	F: TGGCCAGAGGGGCCAGAAGGGCAAA R: CCCCTCTGGGTCTAGAGTTTGCCCT
KRTAP13-1	336H05 **368J2** 408J12 413H10	F: GCAAAGGCTACTTCCTGGTCTA R: ATTGGATGGCAGGATCCACAG	109	F: TCCAGAAGCTGTGGGTCCAGTGG R: TCCAGAACCCAGAGATCCACTGGA
NCOA6	34F15 46J23 **59N23** 86O24	F: CCCAAGATTTTCTAAAGACAGGAA R: CTGGTCAGTATGGGCTTATCTCTT	151	F: CAGCTGTGTTTACAACTCCTCCAGCCAAG R: CTGGTCAGTATGGGCTTATCTCTTGGCTG
SOX9	13O23 30B6 32I21 58P4 68P18 115A15 122H18 169O5 172F10 186L14 202F10 **231H17** 249C14 279H10 297J24 306O5	F: AAATGCTCTTATTTTTCCAACAGC R: AATCACAAAGCCTGAGGAATTAAG	220	F: GTGTTATGGGATCAGTTTGGGGGGTTA R: CTGAGGAATTAAGCAAAGCTAACCCCC
TNFSF12	133N9 169O5 **172F10**	F: GACCTGAATCCCCAGACAGA R: GTGGTTTCCGGCCTTTAGGT	94	F: AGCCAGGACACCGTGTCTTTCCTG R: GAGGCCGAACCAGTTTCAGGAAAG
ZIC1	127I17 **135I16**	F: AGTCCGCGTTCAGAGCACTAT R: GAAAGTTTTGTTGCACGACTTTTT	192	F: GCGCCGGCGCTTTCTTCCGCTACATG R: CTGTTTGATGGGCTGGCGCATGTAGC
ZIC5	211H22 **224A3**	F: GCAAACTTTCTGCAAGTGCAAC R: GGAAGCCTGTCATATTCTGAAAC	199	F: AGGGGGCACGAAGCGAAAGCGAAG R: CTGTGCTCACTGACGCCTTCGCTT

BAC numbers in bold denote those that were used for gene chromosomal localization by FISH.

Based on these predictions, BACs containing new genes were co-hybridized with a differently labeled reference gene from the cytogenetic map (**Table 2**). Biotin- and dig-labeled probes were detected with avidin-FITC (Vector Laboratories) and anti-dig-rhodamine (Roche Applied Science), respectively. Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) and identified according to the nomenclature proposed by Balmus et al. (2007) and Avila et al. (2014b). Images were captured and analyzed using a Zeiss Axioplan 2 fluorescence microscope, equipped with the Isis Version 5.2 (MetaSystems GmbH) software. At least 10 images were captured and analyzed for each experiment.

RESULTS

Altogether, we identified 41 BAC clones that collectively contained the 11 genes of interest. Clones for individual genes were identified by PCR with gene-specific primers (**Table 1**), and one clone per each gene was selected for FISH mapping. In this manner, we assigned 11 BAC clones to eight different alpaca autosomes. Most of the candidate genes were mapped to a specific G-band or a range of G-bands (**Table 2**). Previously mapped reference markers (Avila et al., 2014b) confirmed chromosome identification and helped to position new genes in the centromere-telomere field (**Figure 1**). Four genes were

TABLE 2 | Summary data of newly mapped genes and reference markers (Avila et al., 2014b).

	Gene					Reference marker		
	Symbol	Name	Chromosomal location	VicPac2.0.2 scaffold	CHORI 246 BAC Clone	Name	CHORI 246 BAC Clone	Location
Fiber growth candidate genes	*COL1A1*	Collagen type I alpha 1 chain	16q13	**377**	204B18	*DDX52*	18J7	16p14prox
	CTNNB1	Catenin beta 1	17q12-q13	23	150A21	*MITF*	33H2	17q14
	DAB2IP	Disabled homolog 2-interacting protein	4q34	52	101B6	*GG_478*	71E21	4q34
	KRT15	Keratin 15	16q12-q13	**450**	268A9	*AP2B1*	156N10	16p13
	KRTAP13-1	Keratin Associated Protein 13-1 Like	1q33	101	368J2	*SOX2*	24K2	1q19
	TNFSF12	TNF superfamily member 12	16p13	**387**	172F10	*KCNJ16*	408P6	16q16
Fiber color candidate genes	*NCOA6*	Nuclear receptor coactivator 6	19q12	34	59N23	*ASIP*	18C13	19q12
						BMP7	93P6	19q22
	ZIC1	Zinc finger protein ZIC 1	1q13-q14	**35**	135I16	*SOX2*	24K2	1q18-q21
	ZIC5	Zic family member 5	14q15-q16	**84**	224A3	*RB1*	89N13	14p13
	SOX9	SRY-box 9	16q17	15	231H17	*KCNJ16*	408P6	16q16
	ALX3	ALX homeobox 3	9q24-q25	4	115I10	*GG1068*	2N23	9q14

VicPac2.0.2 scaffolds in bold denote those that were chromosomally assigned first time in this study.

located in chromosome 16 (chr16), and 2 genes in chr1, whereas the remaining five genes mapped to five different chromosomes (**Figure 1** and **Table 2**). In chr19, the *NCOA6* gene overlapped with *ASIP* in 19q12, and their relative order was resolved by interphase FISH using *BMP7* as the second reference marker. The order of the three genes was revealed as cen-*ASIP-NCOA6-BMP7*-tel (**Figure 1F**, far right). Location of *CTNNB1*, *DAB2IP*, and *SOX9* in chr17, chr4 and chr16, respectively, was confirmed by co-hybridized reference markers. No genes were assigned to chromosome arms that previously did not have a mapped marker. No discrepancies of the known conserved synteny blocks between camelids, cattle and human (Balmus et al., 2007) were observed.

DISCUSSION

We report the cytogenetic mapping of 11 new genes in the alpaca genome, which together with prior FISH maps (Avila et al., 2014a,b) takes the tally of all chromosomally mapped markers for this species to 241. This is not a high number but an important step forward for the development of chromosomally anchored reference genomes for the alpaca and other camelids. Furthermore, among the 11 markers mapped in this study, five belong to VicPac2.0.2 scaffolds that were not represented in previous maps (Avila et al., 2014b). This implies that the entire scaffold 35, anchored by *ZIC1*, maps to chr1; scaffold 84, anchored by *ZIC5*, maps to chr14, and scaffolds 377, 387, and 450, anchored by *COL1A1*, *TNFSF12*, and *KRT15*, respectively, map to chr16 (**Table 2**).

As the goal of this study was cytogenetic mapping of candidate genes related to fiber growth and color synthesis, we bioinformatically inspected all VicPac2.0.2 scaffolds containing the 11 mapped markers (**Table 2**) for additional genes of interest. In scaffold 450 (331,325 bp, NW_005883152.1), which was newly assigned to chr16q12-q13 by FISH mapping *KRT15* (**Figure 1G**), there is a tandemly arranged cluster of five more keratin genes around *KRT15*, viz., 5′- *KRT31- KRT15 - KRT13- KRT9- KRT14-KRT16 – 3′* (Fernández et al., 2019). Thus, our results allow the assignment of five more keratin genes to chr16q12-q13 (**Table 3**). This makes alpaca chr16 as a main target for identifying sequence variants potentially associated with hair texture and growth because three of the six candidate genes for fiber growth characteristics (**Table 2**), viz., *KRT15* with the keratin cluster, *COL1A1*, and *TNFSF12* (Fernandez, 2015), map to this chromosome. This also implies that phenotypic characteristics determined by these genes may show particular inheritance patterns due to close linkage. Among the many known molecular components of the mammalian hair follicle (Rompolas and Greco, 2013), keratins and collagens are perhaps most studied (Toivola et al., 2015) and associated with various hair texture characteristics in several mammalian species including humans (Shimomura et al., 2010), dogs (Cadieu et al., 2009), horses (Balmer et al., 2017; Morgenthaler et al., 2017) and alpacas (Fan et al., 2011). Here we considered as candidate genes for alpaca hair texture also genes related to apoptosis regulation and formation of the hair follicle, such as *CTNNB1*, *TNFSF12*, and *DAB2IP*. *TNFSF12* and *DAB2IP* have roles in WNT/β-catenin signaling system (Xie et al., 2010), which controls hair follicle morphogenesis and stem cell differentiation in the skin (Huelsken et al., 2001). SNP variants in these genes have been associated with traits of interest (Farhadian et al., 2018) and used for genomic selection programs in sheep, goat (Rupp et al., 2016) and cattle (Wiggans et al., 2017).

Therefore, microsatellites that have been identified in the alpaca *COL1A1*, *TNFSF12*, and *DAB2IP* (Fernandez, 2015) are potential polymorphic markers for selection in this species.

Among the candidate genes for hair color, mapping *NCOA6* to chr19q12 was of particular interest because it anchored a closely linked group of several other potential coat color genes

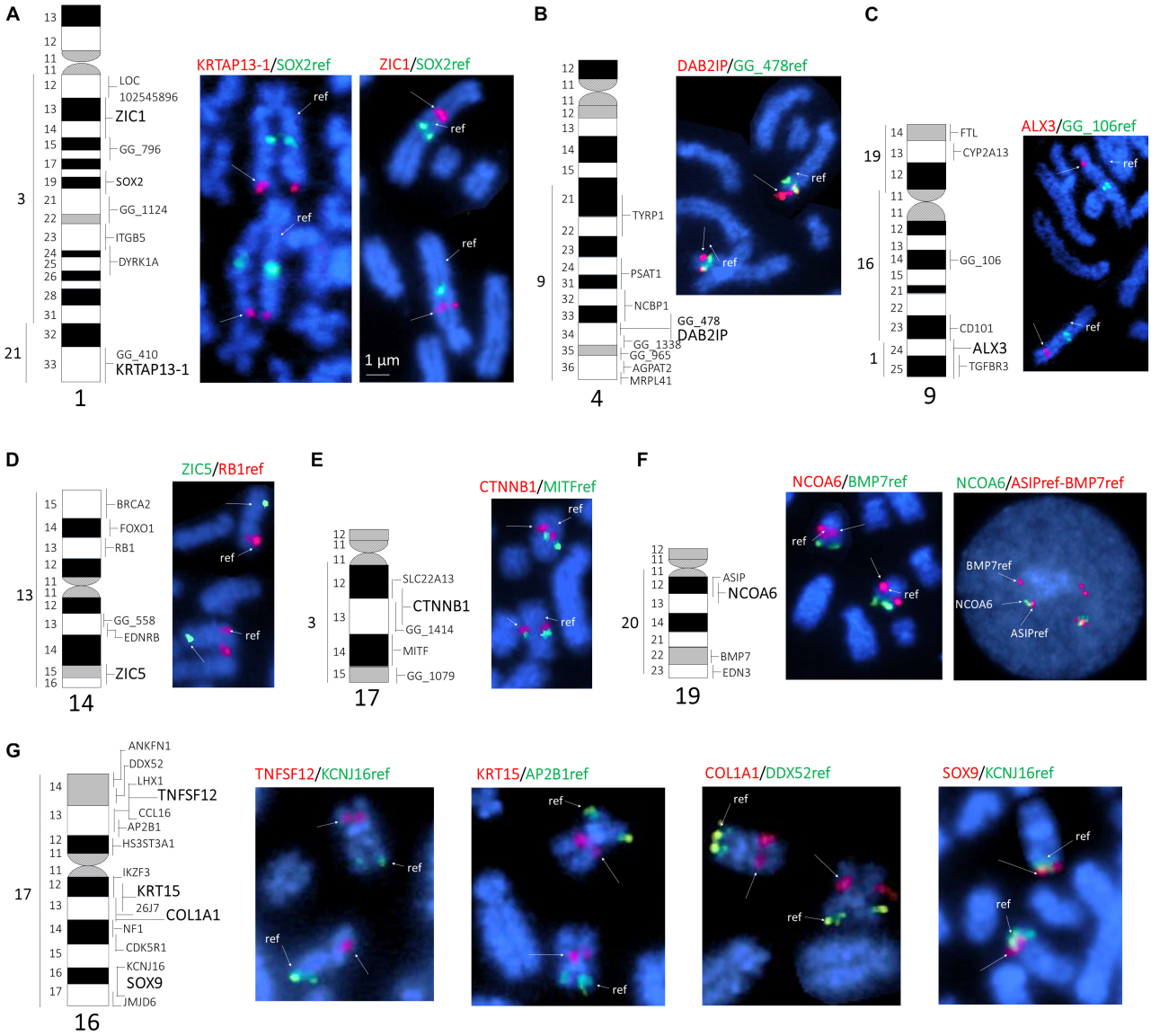


FIGURE 1 | FISH mapping of selected genes to alpaca chromosomes. **(A)** *ZIC1* and *KRTAP13-1* to chr1; **(B)** *DAB2IP* to chr2; **(C)** *ALX3* to chr9; **(D)** *ZIC5* to chr14; **(E)** *CTNNB1* to chr17; **(F)** *NCOA6* to chr16, and **(G)** *TNFSF12*, *KRT15*, *COL1A1*, and *SOX9* to chr16. The left side of each sub-figure includes a G-banded ideogram showing conserved synteny with human chromosomes (far left) and all cytogenetically mapped markers (right). The markers mapped in this study are in larger and bold font. The right side of each sub-figure shows partial metaphase spreads with dual-color FISH results. The newly mapped genes are denoted with arrows; reference genes for chromosome identification are denoted with arrows and "ref." The symbols of newly mapped and reference genes are shown above each figure in green or red font colors, which match with the colors of FISH signals on chromosomes. In **(F)**, we show refined ordering of *NCOA6* in relation to *ASIP* in interphase chromosomes (far right) using *BMP7* is an anchoring marker; the order is cen-*ASIP*-*NCOA6*-*BMP7*-tel.

from scaffold 34 (12,494,946 bp, NW_005882736.1) to this chromosome (**Table 3**). The closely linked gene cluster comprises *RALY*, *EIF2S2*, *XPOTP1*, *ASIP*, *AHCY*, *ITCH*, *PIGU*, *NCOA6*, and *GGT7*, of which only *ASIP* has been previously mapped (Avila et al., 2014b). In this study, we showed that *NCOA6* is overlapping with *ASIP* in chr19q12 (**Figure 1F**) which is consistent with the known organization of the agouti locus in alpacas, where the 5′UTR of the *ASIP* gene contains 142 bp of the *NCOA6* gene sequence (Chandramohan et al., 2013). The role of *ASIP* in regulation of pigment production in mammals is well established (Suzuki, 2013). Mutations in this gene have shown to cause the black coat color phenotype in different species, such as guinea pigs (Lai et al., 2019), black-bone chicken (Yu et al., 2019), sheep (Norris and Whan, 2008; Royo et al., 2008), Iranian Markhoz goats (Nazari-Ghadikolaei et al., 2018), donkeys (Abitbol et al., 2015), horses (Rieder et al., 2001), dogs (Kerns et al., 2004), cats (Eizirik et al., 2003), and impala antelope (Miller et al., 2016). In camelids the agouti signaling protein gene (*ASIP*) is involved in fiber color development in alpacas (Bathrachalam et al., 2011; Chandramohan et al., 2013),

TABLE 3 | Summary data of the genes positionally associated with the genes mapped in this study.

Scaffold VicPac2.0.2	Mapped marker	Positionally associated markers		Inferred VPA chromosomal location
		Gene symbol	Gene name	
450	*KRT15*	*KRT31*	Keratin, type I cuticular Ha1	16q12-q13
		KRT13	Keratin, type I cytoskeletal 13	16q12-q13
		KRT9	Keratin 9	16q12-q13
		KRT14	Keratin 14	16q12-q13
		KRT16	Keratin, type I cytoskeletal 16	16q12-q13
34	*NCOA6*	*RALY*	RALY heterogeneous nuclear ribonucleoprotein	19q12
		EIF2S2	Eukaryotic translation initiation factor 2 subunit beta	19q12
		XPOTP1	Exportin for tRNA pseudogene 1	19q12
		AHCY	Adenosylhomocysteinase	19q12
		ITCH	Itchy E3 ubiquitin protein ligase	19q12
		PIGU	Phosphatidylinositol glycan anchor biosynthesis class U	19q12
		GGT7	Gamma-glutamyltransferase 7	19q12

llamas (Daverio et al., 2016) and dromedaries (Almathen et al., 2018; Alshanbari et al., 2019). Sequence variants (SNPs) in other genes from this linkage group have been associated with color phenotypes in several mammalian species. For example, coat color of the Nanjiang Yellow goat has been associated with SNPs in the *RALY-EIF2S2* locus (Guo et al., 2018), tandem duplication encompassing *ASIP* and *AHCY* coding regions and the *ITCH* promoter region have been reported as the genetic cause of the dominant white coat color of white/tan (A^{Wt}) *agouti* sheep (Norris and Whan, 2008), and *RALY*, *ASIP*, *AHCY*, and *ITCH* are associated with brown and black color coat in Iranian Markohz goat (Nazari-Ghadikolaei et al., 2018). Melanocytes, the cells that are responsible for skin pigmentation, are derived from neural crest cells from all axial levels (Betancur et al., 2010). Therefore, genes involved in neural crest generation, such as *ZIC* genes (Aruga, 2004), are potential candidates for fiber color development. Likewise, *SOX9* is involved in the differentiation of neural crest cells into chondrocytes (Simões-Costa and Bronner, 2015) and cooperates with other cofactors in chondrocytes to regulate expression of *COL2A1* in humans (Hattori et al., 2008). Furthermore, *SOX9* is a key player in ultraviolet B radiation-induced melanocyte differentiation and pigmentation by directly regulating *MITF* (Passeron et al., 2007). *MITF* is involved in melanogenesis regulation in alpaca (Wang et al., 2017) and plays a role in the production of white coat color in the llama (Anello et al., 2019). Finally, *ALX3* is involved in color differentiation in striped rodents (Cuthill et al., 2017), and proposed as a target melanoma gene fusion in humans (Berger et al., 2010). Also, Marín et al. (2018) used the genetic variation of *MC1R* and *ASIP* genes, that control coat color, to differentiate between wild and domestic South American camelids.

In summary, the findings of this study facilitate the improvement and chromosomal assignment of the alpaca genome reference sequence. This, in turn, is critical for correct assembly of newly sequenced individual animals and the discovery of sequence variants in candidate genes for fiber characteristics, coat color and other traits of interest. For instance, Alshanbari et al. (2019) have recently assign the *MC1R* gene to camelid chr21 that is not in line with the human-camelids Zoo-FISH synteny map. In addition, improving the alpaca cytogenetic map provides new molecular markers for clinical cytogenetics in alpacas and other camelids, thus facilitating chromosome identification in these complex karyotypes. Finally, cytogenetic mapping of specific genes refines the Zoo-FISH information (Balmus et al., 2007), reveals new evolutionary conserved synteny segments between camelids and other mammals, and adds to our knowledge about camelid chromosome evolution.

ETHICS STATEMENT

The cell cultures were prepared from alpaca blood samples obtained in accordance with the United States Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training, approved by Animal Use Protocol AUP #2011-96, # 2018-0342 CA and CRRC #09-47 at Texas A&M University.

AUTHOR CONTRIBUTIONS

FPdL and GG conceived and supervised the study. MM conducted the experimental work. FA contributed to the BAC screening. MM and TR analyzed the data. MM wrote the manuscript in close consultation with FPdL, TR, and GG. All authors read and approved the final version of the manuscript.

FUNDING

The authors acknowledge the financial support from CONCYTEC through project 125-2015 FONDECYT, and VLIR-UOS through funding of the UNALM (IUC) programme.

ACKNOWLEDGMENTS

Opinions of the author(s) do not automatically reflect those of either the Belgian Government or VLIR-UOS, and can neither bind the Belgian Government nor VLIR-UOS.

REFERENCES

Abitbol, M., Legrand, R., and Tiret, L. (2015). A missense mutation in the agouti signaling protein gene (ASIP) is associated with the no light points coat phenotype in donkeys. *Genet. Sel. Evol.* 47:28. doi: 10.1186/s12711-015-0112-x

Almathen, F., Elbir, H., Bahbahani, H., Mwacharo, J., and Hanotte, J. (2018). Polymorphisms in MC1R and ASIP genes are associated with coat color variation in the arabian camel. *J. Hered.* 109, 700–706. doi: 10.1093/jhered/esy024

Alshanbari, F., Castaneda, C., Juras, R., Hillhouse, A., Mendoza, M. N., Gutiérrez, G. A., et al. (2019). Comparative FISH-Mapping of MC1R, ASIP, and TYRP1 in new and old world camelids and association analysis with coat color phenotypes in the dromedary (*Camelus dromedarius*). *Front. Genet.* 10:340. doi: 10.3389/fgene.2019.00340

Anello, M., Daverio, M. S., Silbestro, M. B., Vidal-Rioja, L., and Di Rocco, F. (2019). Characterization and expression analysis of KIT and MITF-M genes in llamas and their relation to white coat color. *Anim. Genet.* 50, 143–149. doi: 10.1111/age.12769

Aruga, J. (2004). The role of Zic genes in neural development. *Mol. Cell. Neurosci.* 2004, 205–221.

Avila, F., Baily, M., Perelman, P., Das, P., Pontius, J., Chowdhary, R., et al. (2014a). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*Lama pacos*). *Cytogenet. Genome. Res.* 144, 196–207. doi: 10.1159/000370329

Avila, F., Das, P., Kutzler, M., Owens, E., Perelman, P., Rubes, J., et al. (2014b). Development and application of camelid molecular cytogenetic tools. *J. Hered.* 2014, 858–869. doi: 10.1093/jhered/ess067

Avila, F., Baily, M. P., Merriwether, D. A., Trifonov, V. A., Rubes, J., Kutzler, M. A., et al. (2015). A cytogenetic and comparative map of camelid chromosome 36 and the minute in alpacas. *Chromo. Res.* 23, 237–251. doi: 10.1007/s10577-014-9463-3

Balmer, P., Bauer, A., Pujar, S., McGarvey, K. M., Welle, M., Galichet, A., et al. (2017). A curated catalog of canine and equine keratin genes. *PLoS One* 12:e0180359. doi: 10.1371/journal.pone.0180359

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'brien, P. C., Alkalaeva, E. S., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative *Cetartiodactyla* ancestral karyotype. *Chromo. Res.* 15, 499–515.

Bathrachalam, C., La Manna, V., Renieri, C., and La Terza, A. (2011). "Asip and MC1R cDNA polymorphism in alpaca," in *Fibre Production in South American Camelids and Other Fibre Animals, Sprinter Book*, eds M. Á Pérez, J. P. Gutiérrez, I. Cervantes, and M. J. Alcalde (Amsterdam: Wageningen Academic Publishers), 93–96.

Berger, M. F., Levin, J. Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., et al. (2010). Integrative analysis of the melanoma transcriptome. *Genome Res.* 2010 Apr 20, 413–427. doi: 10.1101/gr.103697.109

Betancur, P., Bronner-Fraser, M., and Sauka-Spengler, T. (2010). Assembling neural crest regulatory circuits into a gene regulatory network. *Annu. Rev. Cell Dev. Biol.* 2010, 581–603. doi: 10.1146/annurev.cellbio.042308.113245

Bianchi, N. O., Larramendy, M. L., Bianchi, M. S., and Cortes, L. (1986). Karyological conservation in south american camelids. *Experientia* 42, 622–624.

Cadieu, E., Neff, M. W., Quignon, P., Walsh, K., Chase, K., Parker, H. G., et al. (2009). Coat variation in the domestic dog is governed by variants in three genes. *Science* 326, 150–153. doi: 10.1126/science.1177808

Chandramohan, B., Renieri, C., La Manna, V., and La Terza, A. (2013). The alpaca agouti gene: genomic locus, transcripts and causative mutations of eumelanic and pheomelanic coat color. *Gene* 521, 303–310. doi: 10.1016/j.gene.2013.03.060

Crispín, M. (2008). *Productividad y distribución de fibra de alpaca en la región de Huancavelica: Un análisis comparativo entre Huancavelica y Puno*. [dissertation/College degree]. Lima: Universidad Nacional Mayor de San Marcos.

Cruz, A., Cervantes, I., Burgos, A., Morante, R., and Gutierrez, J. P. (2017). Genetic parameters estimation for preweaning traits and their relationship with reproductive, productive and morphological traits in alpaca. *Animal* 11, 746–754. doi: 10.1017/S175173111600210X

Cuthill, I. C., Allen, W. L., Arbuckle, K., Caspers, B., Chaplin, G., Hauber, M. E., et al. (2017). The biology of color. *Science* 357:470.

Daverio, M. S., Rigalt, F., Romeroc, S., Vidal-Riojaa, L., and Di Rocco, F. (2016). Polymorphisms in MC1R and ASIP genes and their association with coat color phenotypes in llamas (*Lama glama*). *Small Rumin. Res.* 144, 83–89. doi: 10.1016/j.smallrumres.2016.08.003

Di Berardino, D., Nicodemo, D., Coppola, G., King, A. W., Ramunno, L., Cosenza, G. F., et al. (2006). Cytogenetic characterization of alpaca (*Lama pacos, fam. Camelidae*) prometaphase chromosomes. *Cytogenet. Genome Res.* 115, 138–144.

Eizirik, E., Yuhki, N., Johnson, W. E., Menotti-Raymond, M., Hannah, S. S., and O'Brien, S. J. (2003). Molecular genetics and evolution of melanism in the cat family. *Curr. Biol.* 13, 448–453. doi: 10.1016/S0960-9822(03)128-123

Fan, R., Dong, Y., Cao, J., Bai, R., Zhu, Z., Li, P., et al. (2011). Gene expression profile in white alpaca (*Vicugna pacos*) skin. *Animal* 5, 1157–1161. doi: 10.1017/S1751731111000280

Farhadian, M., Rafat, S. A., Hasanpur, K., Ebrahimi, M., and Ebrahimie, E. (2018). Cross-species meta-analysis of transcriptomic data in combination with supervised machine learning models identifies the common gene signature of lactation process. *Front. Genet.* 9:235. doi: 10.3389/fgene.2018.00235

Fernández, A. G., Gutiérrez, G. A., and Ponce de León, F. A. (2019). Bioinformatic identification of single nucleotide polymorphisms (SNPs) in candidate genes for fiber characteristics in alpacas (*Vicugna pacos*). *Revista peruana de biología* 26, 087–094. doi: 10.15381/rpb.v26i1.15911

Fernandez, D. (2015). *Búsqueda de genes relacionados a la síntesis de la fibra y marcadores SSR en los ESTs de piel de alpaca Vicugna pacos*. [dissertation/College degree]. Lima: Universidad Nacional Mayor de San Marcos.

Florez, F. (2016). *Caracterización de marcadores genéticos en genes que codifican a proteínas asociadas a queratina y evaluación de la asociación del gen KRTAP11-1 al diámetro de fibra en alpaca (Vicugna pacos)* siguiendo una aproximación de gen candidato. [dissertation/Master's thesis]. Lima: Universidad Peruana Cayetano Heredia.

Guo, J., Tao, H., Li, P., Li, L., Zhong, T., Wang, L., et al. (2018). Whole-genome sequencing reveals selection signatures associated with important traits in six goat breeds. *Sci Rep.* 8:10405. doi: 10.1038/s41598-018-28719-w

Hack, W. (2001). *The Peruvian Alpaca Meat and Hide Industries RIRDC Publication No. 01/19. Project No. TA001-18*. https://www.agrifutures.com.au/wp-content/uploads/publications/01-019.pdf (accessed March 2011).

Hattori, T., Coustry, F., Stephens, S., Eberspaecher, H., Takigawa, M., Yasuda, H., et al. (2008). Transcriptional regulation of chondrogenesis by coactivator Tip60 via chromatin association with Sox9 and Sox5. *Nucleic Acids Res.* 36, 3011–3024. doi: 10.1093/nar/gkn150

Hsu, T. C., and Benirschke, K. (1967). *An Atlas of Mammalian Chromosomes*. New York, NY: Springer.

Huelsken, J., Vogel, R., Erdmann, B., Cotsarelis, G., and Birchmeier, W. (2001). beta-Catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell* 18, 533–545.

Kerns, J. A., Newton, J., Berryere, T. G., Rubin, E. M., Cheng, J. F., Schmutz, S. M., et al. (2004). Characterization of the dog agouti gene and a non agouti mutation in German shepherd dogs. *Mamm. Genome* 15, 798–808. doi: 10.1007/s00335-004-2377-1

Lai, W., Hu, M., Zhu, W., Yu, F., Bai, C., Zhang, J., et al. (2019). A 4-bp deletion in the ASIP gene is associated with the recessive black coat colour in domestic guinea pigs (*Cavia porcellus*). *Anim. Genet.* 50, 190–191. doi: 10.1111/age.12766

Marín, J. C., Rivera, R., Varas, V., Cortés, J., Agapito, A., Chero, A., et al. (2018). Genetic variation in coat colour genes MC1R and ASIP provides insights into domestication and management of south american camelids. *Front. Genet.* 9:487. doi: 10.3389/fgene.2018.00487

Miller, S. M., Guthrie, A. J., and Harper, C. K. (2016). Single base-pair deletion in ASIP exon 3 associated with recessive black phenotype in impala (*Aepyceros melampus*). *Anim. Genet.* 47, 511–512. doi: 10.1111/age.12430

Morante, R., Goyache, F., Burgos, A., Cervantes, I., Pérez-Cabal, M. A., and Gutiérrez, J. P. (2009). Genetic improvement for alpaca fibre production in the peruvian altiplano: the pacomarca experience. *Anim. Genetic Res. Infor.* 2009, 37–43. doi: 10.1017/S1014233909990307

Morgenthaler, C., Diribarne, M., Capitan, A., Legendre, R., Saintilan, R., Gilles, M., et al. (2017). A missense variant in the coil1A domain of the keratin 25 gene is associated with the dominant curly hair coat trait (Crd) in horse. *Genet. Sel. Evol.* 49:85. doi: 10.1186/s12711-017-0359-355

Nazari-Ghadikolaei, A., Mehrabani-Yeganeh, H., Miarei-Aashtiani, S. R., Staiger, E. A., Rashidi, A., and Huson, H. J. (2018). Genome-wide association studies identify candidate genes for coat color and mohair traits in the Iranian markhoz goat. *Front. Genet.* 9:105. doi: 10.3389/fgene.2018.00105

Norris, B. J., and Whan, V. A. (2008). A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Res.* 18, 1282–1293. doi: 10.1101/gr.072090.107

Passeron, T., Valencia, J. C., Bertolotto, C., Hoashi, T., Le Pape, E., Takahashi, K., et al. (2007). SOX9 is a key player in ultraviolet B-induced melanocyte differentiation and pigmentation. *Proc. Natl. Acad. Sci. U.S.A.* 28, 13984–13989.

Quispe, E. C., Rodríguez, T. C., Iñiguez, L. R., and Mueller, J. P. (2009). Producción de fibra de alpaca, llama, vicuña y guanaco en Sudamérica. *Anim. Genetic Res. Inform.* 45, 1–14. doi: 10.1017/S1014233909990277

Raudsepp, T., and Chowdhary, B. P. (2008). FISH for mapping single copy genes. *Methods Mol. Biol.* 422, 31–49. doi: 10.1007/978-1-59745-581-7_3

Rieder, S., Taourit, S., Mariat, D., Langlois, B., and Guerin, G. (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (*Equus caballus*). *Mamm.Genome* 12, 450–455. doi: 10.1007/s003350020017

Rompolas, P., and Greco, V. (2013). Stem cell dynamics in the hair follicle niche. *Semin. Cell Dev. Biol.* 2, 34–42.

Royo, L. J., Alvarez, I., Arranz, J. J., Fernandez, I., Rodriguez, A., Perez-Pardal, L., et al. (2008). Differences in the expression of the ASIP gene are involved in the recessive black coat colour pattern in sheep: evidence from the rare Xalda sheep breed. *Anim. Genet.* 39, 290–293. doi: 10.1111/j.1365-2052.2008.01712.x

Rupp, R., Mucha, S., Larroque, H., McEwan, J., and Conington, J. (2016). Genomic application in sheep and goat breeding. *Anim Front.* 2016, 39–44.

Shimomura, Y., Wajid, M., Petukhova, L., Kurban, M., and Christiano, A. M. (2010). Autosomal-dominant woolly hair resulting from disruption of keratin 74 (KRT74), a potential determinant of human hair texture. *Am. J. Hum. Genet.* 86, 632–638. doi: 10.1016/j.ajhg.2010.02.025

Simoes-Costa, M., and Bronner, M. E. (2013). Insights into neural crest development and evolution from genomic analysis. *Genome Res.* 2013, 1069–1080. doi: 10.1101/gr.157586.113

Simões-Costa, M., and Bronner, M. E. (2015). Establishing neural crest identity: a gene regulatory recipe. *Development* 142, 242–257. doi: 10.1242/dev.105445

Suzuki, H. (2013). Evolutionary and phylogeographic views on Mc1r and asip variation in mammals. *Genes Genet. Syst.* 2013, 155–164.

Taylor, K. M., Hungerford, D. A., Snyder, R. L., and Ulmer, F. A. Jr. (1968). Uniformity of karyotypes in the camelidae. *Cytogenetics* 7, 8–15.

Toivola, D. M., Boor, P., Alam, C., and Strnad, P. (2015). Keratins in health and disease. *Curr. Opin. Cell Biol.* 32, 73–81. doi: 10.1016/j.ceb.2014.12.008

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3 - new capabilities and interfaces. *Nucleic Acids Res.* 40:e115.

Wang, R., Chen, T., Zhao, B., Fan, R., Ji, K., Yu, X., et al. (2017). FGF21 regulates melanogenesis in alpaca melanocytes via ERK1/2-mediated MITF downregulation. *Biochem. Biophys. Res. Commun.* 490, 466–471. doi: 10.1016/j.bbrc.2017.06.064

Wiggans, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: the USDA Experience. *Annu. Rev. Anim. Biosci.* 5, 309–327. doi: 10.1146/annurev-animal-021815-111422

Xie, D., Gore, C., Liu, J., Pong, R. C., Mason, R., Hao, G., et al. (2010). Role of DAB2IP in modulating epithelial-to-mesenchymal transition and prostate cancer metastasis. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2485–2490. doi: 10.1073/pnas.0908133107

Yu, S., Wang, G., and Liao, J. (2019). Association of a novel SNP in the ASIP gene with skin color in black-bone chicken. *Anim. Genet.* 50, 283–286. doi: 10.1111/age.12768

Yucra, L. E. (2017). *Sistema de comercialización y situación sociocultural, económica y ambiental de la cadena de producción de la fibra de alpaca en el distrito de Macusani, provincia de Carabaya, Puno*. [dissertation/Master's thesis]. Lima: Pontificia Universidad Católica del Perú.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Mendoza, Raudsepp, Alshanbari, Gutiérrez and Ponce de León. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 June 2019
doi: 10.3389/fgene.2019.00586

Chromosome-Level Alpaca Reference Genome *VicPac3.1* Improves Genomic Insight Into the Biology of New World Camelids

*Mark F. Richardson[1,2†], Kylie Munyard[3‡], Larry J. Croft[1], Theodore R. Allnutt[4], Felicity Jackling[5], Fahad Alshanbari[6§], Matthew Jevit[6], Gus A. Wright[6], Rhys Cransberg[3], Ahmed Tibary[7‖], Polina Perelman[8¶], Belinda Appleton[2] and Terje Raudsepp[6]**

[1] Genomics Centre, Deakin University, Geelong, VIC, Australia, [2] Centre for Integrative Ecology, Deakin University, Geelong, VIC, Australia, [3] School of Pharmacy and Biomedical Sciences, Curtin Health Innovation Research Institute, Curtin University, Perth, WA, Australia, [4] Bioinformatics Core Research Group, Deakin University, Geelong, VIC, Australia, [5] Department of Genetics, The University of Melbourne, Melbourne, VIC, Australia, [6] Department of Veterinary Pathobiology, Texas A&M University, College Station, TX, United States, [7] Center for Reproductive Biology, Washington State University, Pullman, WA, United States, [8] Institute of Molecular and Cellular Biology, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia

The development of high-quality chromosomally assigned reference genomes constitutes a key feature for understanding genome architecture of a species and is critical for the discovery of the genetic blueprints of traits of biological significance. South American camelids serve people in extreme environments and are important fiber and companion animals worldwide. Despite this, the alpaca reference genome lags far behind those available for other domestic species. Here we produced a chromosome-level improved reference assembly for the alpaca genome using the DNA of the same female Huacaya alpaca as in previous assemblies. We generated 190X Illumina short-read, 8X Pacific Biosciences long-read and 60X Dovetail Chicago® chromatin interaction scaffolding data for the assembly, used testis and skin RNAseq data for annotation, and cytogenetic map data for chromosomal assignments. The new assembly *VicPac3.1* contains 90% of the alpaca genome in just 103 scaffolds and 76% of all scaffolds are mapped to the 36 pairs of the alpaca autosomes and the X chromosome. Preliminary annotation of the assembly predicted 22,462 coding genes and 29,337 isoforms. Comparative analysis of selected regions of the alpaca genome, such as the major histocompatibility complex (MHC), the region involved in the *Minute Chromosome Syndrome* (MCS) and candidate genes for high-altitude adaptations, reveal unique features of the alpaca genome. The alpaca reference genome *VicPac3.1* presents a significant improvement in completeness, contiguity and accuracy over *VicPac2* and is an important tool for the advancement of genomics research in all New World camelids.

Keywords: alpaca, reference genome, *VicPac3.1*, chromosome-level, Dovetail Chicago, MHC, *Minute Chromosome Syndrome*, high-altitude adaptations

OPEN ACCESS

Edited by:
Pamela Burger,
University of Veterinary Medicine, Austria

Reviewed by:
Huiguang Wu,
Yangzhou University, China
Maria V. Sharakhova,
Virginia Tech, United States

****Correspondence:***
Terje Raudsepp
traudsepp@cvm.tamu.edu
orcid.org/0000-0003-2276-475X
†orcid.org/0000-0002-1650-0064
‡orcid.org/0000-0002-5113-8646
§orcid.org/0000-0002-4674-1360
‖orcid.org/0000-0003-1546-3342
¶orcid.org/0000-0002-0982-5100

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 January 2019*
Accepted: *04 June 2019*
Published: *21 June 2019*

Citation:
Richardson MF, Munyard K, Croft LJ, Allnutt TR, Jackling F, Alshanbari F, Jevit M, Wright GA, Cransberg R, Tibary A, Perelman P, Appleton B and Raudsepp T (2019) Chromosome-Level Alpaca Reference Genome VicPac3.1 Improves Genomic Insight Into the Biology of New World Camelids. Front. Genet. 10:586. doi: 10.3389/fgene.2019.00586

INTRODUCTION

Alpacas and llamas were domesticated in the high Andes around 9,000 years ago and have been associated with humans for as long as cattle, horses and dogs (Wheeler, 1995; Bruford et al., 2003). It is thought that the ancient Incan civilization owed success largely to llama dung, which provided fertilizer and enabled corn to be cultivated at very high altitudes. Today, alpacas continue to serve the rural families of the Altiplano as an important source of fiber and meat (Cruz et al., 2017). In addition, alpacas are also gaining popularity worldwide, mainly for their high quality fiber, and as a docile companion species. In addition, alpacas and camelids in general, are species of broader interest for several fields in biology and biomedical sciences. For example, the family Camelidae forms the most basal clade in the phylogeny of the eutherian order Cetartiodactyla (Murphy et al., 2005; Zhou et al., 2011) and is, thus, a key-group in the mammalian evolutionary tree, and is being used to aid in the annotation of the human genome (Genome 10K Community of Scientists, 2009). Further, genetic relationships between South American camelids, the domesticated alpaca (*Vicugna pacos*) and llama (*Lama glama*), and the wild guanaco (*Lama guanicoe*) and vicuña (*Vicugna vicugna*), are intriguing and still not completely resolved (Bruford et al., 2003; Barreta et al., 2013; Marin et al., 2018). All camelids are uniquely adapted to extreme environments – the New World species to high altitude and the Old World camels to arid desert environments (Wu et al., 2014), due to these adaptations their genomes may reveal important signatures of natural or human selection. Camelids are also of biomedical interest because of the presence of small and functionally efficient heavy chain-only antibodies, which are not found in other mammalian groups (Flajnik et al., 2011; Griffin et al., 2014; Cohen, 2018).

Despite being a species of broad interest, the analysis of camelid genomes, including that of the alpaca, had a late start and lags behind other domesticated species. Camelid karyotypes were described in the 1980s (Bianchi et al., 1986), showing that all extant species have a conserved diploid number (2n = 74) and very similar chromosome morphology. Yet, the first cytogenetic and comparative chromosome maps for these species emerged only recently (Balmus et al., 2007; Avila et al., 2014a,b, 2015), almost concurrently with genome sequencing projects. At present, there are two annotated sequence assemblies for the alpaca that are available at all main Genome Browsers such as NCBI[1], UCSC[2] and Ensembl[3]: *VicPac1* (version 1.0) and *VicPac2* (version 2.0.1). Both used DNA from the same female Huacaya individual. The first assembly was generated at the Broad Institute by Sanger sequencing and has 2.51X genome coverage, the second was assembled at Washington University by combining the former Sanger reads with newly generated 454 GS FLX data. This resulted in an assembly with 22X genome coverage and annotation for 24,553 genes and 33,208 proteins. *VicPac1* and *VicPac2* form the alpaca reference genome and are currently the main tools for alpaca genomics. There is also a third assembly, *Vipacos_V1.0*, which was generated for the comparison of genomic signatures of selection and adaptations between the dromedary, Bactrian camel and alpaca (Wu et al., 2014). *Vipacos_V1.0* was assembled from short-read Illumina data and reached 72.5X genome coverage, but is not integrated with *VicPac1* or *VicPac2*. Despite this progress, all three alpaca assemblies are relatively short – 2 billion DNA base-pairs (2 Gb) instead of the anticipated 2.5–3 Gb; all are fragmented into a large number of contigs and scaffolds, and none have scaffolds assigned to chromosomes. The overall utility of these datasets as an alpaca reference genome to serve the interests of researchers, breeders and the health and welfare of the animals, is therefore limited and needs improvement.

The aim of this study was to re-sequence, re-assemble *de novo* and re-annotate the alpaca genome using the same female Huacaya DNA donor as in *VicPac1* and *VicPac2*. We used next generation long- and short-read sequencing platforms to generate the data and initial assembly; Dovetail Chicago® scaffolding and HiRise™ for advanced assembly; RNAseq and bioinformatics pipelines for annotation, and cytogenetic comparative map data to anchor sequence scaffolds to chromosomes.

RESULTS AND DISCUSSION

Genome and Assembly Features

The genome of a female Huacaya alpaca was sequenced generating ∼190X genome coverage of paired-end (PE) and mate-pair (MP) short-read Illumina data (2.72 billion PE reads, 272 Gb; 1.52 billion MP reads, 152 Gb), ∼8X genome coverage of Pacific Biosciences (PacBio) long-read data (2.4 million subreads; 18.0 Gb), and ∼60X genome coverage Dovetail Chicago® chromatin interaction scaffolding data (459 million PE reads; 137.7 Gb). A multi-stage assembly improvement strategy was applied through four separate assembly iterations. Firstly, we produced a hybrid *de novo* assembly using the PE and MP short-read data together with the Sanger and 454 data from the *VicPac1* and *VicPac2* assemblies, respectively. This assembly (*Qmas1*) had more contigs and scaffolds than *VicPac2*, lower scaffold N50, but higher contig N50 (**Table 1**). Next, we integrated *Qmas1* and *VicPac2* to produce a meta-assembly (*Qmas1*/*VicPac2*) that resulted in contiguity improvements, namely a reduction in the number of contigs and scaffolds and the simultaneous increase in contig and scaffold N50s (**Table 1**). The next iteration of the assembly incorporated the ∼ 8X PacBio long-read data and resulted in modest improvements to the assembly (designated *VicPac3*) compared to previous iterations (**Table 1**). The final assembly iteration involved scaffolding the *VicPac3* assembly with the MP short read data and Dovetail Chicago® data. This final assembly also resulted in significant improvements in the assembly metrics (see **Table 1**), including a significant increase of scaffold N50 from 9.86 Mb in *VicPac3* to 24 Mb in *VicPac3.1*. Compared to all previous assemblies, *VicPac3.1* has the best assembly metrics and most importantly, 90% of the assembly sequence length (L90) is contained in just 103 scaffolds (0.1% of all scaffolds; **Table 1**). The remaining 10% of the assembly

[1]https://www.ncbi.nlm.nih.gov/

[2]https://genome.ucsc.edu/

[3]http://www.ensembl.org/index.html

TABLE 1 | Comparative summary statistics of alpaca genome assemblies.

	VicPac3.1	***VicPac3.0***	***Qmas1/VicPac2***	***Qmas1***	***VicPac2.0***	***VicPac1.0***	***Vipacos_V1.0***
Breed	Huacaya	Huacaya	Huacaya	Huacaya	Huacaya	Huacaya	Huacaya
Sex	Female	Female	Female	Female	Female	Female	Female
Individual	*Carlotta*	*Carlotta*	*Carlotta*	*Carlotta*	*Carlotta*	*Carlotta*	n/a
Assembly size (Gb)	2.12	2.12	2.12	2.66	2.17	2.96	2.01
Contig N50 (kb)	35.72	35.75	35.75	306.09	29.07	3.91	66.3
Number of contigs	204,817	204,577	205,666	719,860	412,904	721,292	75,733
Scaffold N50 (Mb)	24.02	9.86	9.06	5.83	7.26	0.23	5.1
Scaffold L50	25	64	69	126	86	2,595	–
Number of scaffolds	77,390	78,963	82,481	678,087	276,726	298,413	4,322
Longest scaffold (Mb)	121.37	38.36	38.36	25.07	38.45	5.51	–
GC %	41.4	41.4	41.4	41.6	41.4	39.7	41.5
N's %	4.17	4.09	3.98	2.44	4.31	35.09	–
Repeat %	33.48	–	–	–	34.74		32.1
Reference	This study	This study	This study	This study	NCBI[a], UCSC[b], Ensembl[c]	NCBI, UCSC, Ensembl	Wu et al., 2014

Source: [a]https://www.ncbi.nlm.nih.gov/, [b]https://genome.ucsc.edu/, [c]http://www.ensembl.org/index.html.

sequence length is made up of smaller, fragmented scaffolds. Addition of higher coverage long-read data, for example 20X, compared to the 8X we used, may be needed to generate further improvements to the assembly, through filling gaps and joining scaffolds. The most critical improvements in the contiguity and accuracy of the assembly occurred during the meta-assembly of *Qmas1* and *VicPac2*, and subsequent HiRise™ scaffolding of *VicPac3*. The latter corrected 240 inaccurate assemblies, joined 1813 scaffolds, and essentially improved the size of scaffold N50 and reduced L50 and the total number of scaffolds (**Table 1**).

The GC-content of the alpaca genome was ∼41% and remained the same across all our assembly iterations, and is similar to that reported in prior alpaca assemblies (**Table 1**). The 2.12 Gb size of the re-assembled genome *VicPac3.1* is similar to previous assemblies of the same individual (**Table 1**) but smaller than the 2.63 Gb estimation by k-mer analysis (Wu et al., 2014). Genome size estimation using a range of k-mer frequencies obtained from our short-read data produced size estimates ranging from 2.05 to 2.29 Gb (**Supplementary Figure 1** and **Supplementary Table 1**), which are very similar to the obtained genome sizes for all assemblies in **Table 1** for the same animal, but smaller than the prior k-mer estimation (Wu et al., 2014). On the other hand, measurement of the genome size by flow cytometry using alpaca fibroblasts suggested size of 2.88 Gb with a range of 2.73–3.01 Gb (95% confidence interval; **Supplementary Figure 2**), thus larger than the bioinformatic estimates by us or others. However, it must be noted that the available computational and empirical methods for estimating genome size are subject to very large errors. Furthermore, genome size will vary between individuals. These factors combined may account for the differences between the estimates, and the exact size of the alpaca genome is yet to be determined by additional studies.

The Benchmarking Universal Single-Copy Orthologs (BUSCO)[4] mammalian gene set with 4,104 conserved mammalian orthologs (hereafter BUSCOs) was used to assess genome completeness in terms of recovery of these BUSCOs, to evaluate assembly iterations and compare them to previous alpaca assembly versions. While BUSCO analysis is more appropriate for direct comparison of different genome assemblies within a species, it can provide useful benchmarks when compared to assemblies of other species. Therefore, we also produced BUSCO assessment data for cow, sheep, dromedary and Bactrian camel. Serial improvements in BUSCO scores were observed throughout the iterative assembly process (**Table 2**), with the final assembly, *VicPac3.1*, having the highest BUSCO completeness at 96.1% with 3,944 genes and the lowest number of missing BUSCOs (77 genes; 1.9%). Compared to other available camelid genomes, the final assembly demonstrated comparable, but slightly superior scores across all metrics, suggesting that this assembly is one of the most complete available for camelids, and has completeness scores comparable with the cattle and sheep genomes. The datasets are available in BioProject ID PRJNA544883.

Chromosomal Assignment

Sequences from the available alpaca cytogenetic map (Avila et al., 2014a,b, 2015) and comparative data with human, cattle and pig genomes (Balmus et al., 2007) were used to anchor the alpaca genome sequence assembly to physical chromosomes. In *VicPac3.1*, 75.9% of sequence scaffolds (in bp; ∼1.6 Gb) are mapped to the 36 pairs of alpaca autosomes and the X chromosome (**Table 3** and **Supplementary Table 2**) providing the first chromosome-level assembly for the alpaca, or any camelid genome. Notably, this is a 31.9% increase in the amount of anchored sequence compared to our anchoring of *VicPac2* (44% of sequence scaffolds in bp; 0.96 Gb). Among the most notable improvements are assemblies of 14 alpaca chromosomes, *viz.,* chrs2, 5, 7, 8, 10, 17, 19, 22, 24, 27, 28, 31, 33, and 34 that uniquely correspond to a single

[4]https://busco.ezlab.org/

TABLE 2 | BUSCO analysis of genome completeness.

	Complete and single copy (%)	Complete and duplicated (%)	Fragmented (%)	Missing (%)
Alpaca genomes				
VicPac3.1	96.1	0.7	2.0	1.9
VicPac3.0	94.7	0.7	2.4	2.2
Qmas/VicPac2	95.0	0.7	2.1	2.2
Qmas1	94.2	0.8	2.1	2.9
VicPac2.0.2	93.9	0.8	2.6	2.7
Other camelid genomes				
Camelus dromedarius	95.0	0.5	2.5	2.0
C. bactrianus ferus	94.5	1.2	2.6	1.7
C. bactrianus	95.2	0.5	2.3	2.0
Select mammalian genomes				
Bos taurus	92.4	1.2	3.0	3.4
Ovis aries	92.1	1.1	3.4	3.4

large scaffold (**Table 3**); *VicPac2* only contains 2 chromosomes made up of single scaffolds (chrs19 and 31; **Table 3** and **Supplementary Table 3**). Additionally, the total number of chromosomally anchored scaffolds was reduced from 129 in *VicPac2* to 88 in *VicPac3.1*, while simultaneously increasing the percentage of the genome anchored, further highlighting the significant improvements in contiguity of *VicPac3.1*. Currently, the most contiguous and largest is the 121 Mb scaffold of chr2, which likely represents the entire chromosome. In contrast, chr11 and chr16 remain rather fragmented and correspond to six different scaffolds each. It is notable that three scaffolds, with a total size of 5 Mb, correspond to the smallest autosome, chr36, because no sequences were assigned to this chromosome by Zoo-FISH (Balmus et al., 2007). Despite this progress, assemblies of several large chromosomes remain fragmented and incomplete. For example, eight unique scaffolds correspond to chrX, but cover collectively less than 40 Mb of the anticipated 150 Mb, which is the size of an average mammalian X chromosome[5,6,7].

Genome Annotation

Altogether, preliminary annotation predicted 42,389 genes in the alpaca genome. Of these, 22,462 were coding genes with an average of 14.7 exons per gene. Single-exon genes accounted for 11% (2,519) of these, thus multi-exon genes had an average of 16.5 exons. Overall, the predicted protein coding genes represented 39.6% of assembled sequence with coding exons covering 2.4% of assembled sequence. Coding genes contained 1.3 isoforms on average with a total of 29,337 coding isoforms. The number of genes predicted in the alpaca genome was higher than expected for a vertebrate or mammalian genome (Pennisi, 2012). This may be due to the limited transcriptome depth used to stitch exons into contiguous genes. The human transcriptome now has many terabases of sequencing depth from multiple tissues and conditions to generate a comprehensive (but still incomplete) transcriptome (Steward et al., 2017). The total number of predicted exons in the alpaca is also larger than more comprehensively annotated mammalian genomes, owing to small exonic fragments which may actually be exon extension and truncation events rather than unconnected exons. However, 22,462 predicted coding genes (*e*-value < 1e-20) are similar to the number of known proteins in the mammalian RefSeq database (Pruitt et al., 2014)[8]. It is possible that the remaining 19,927 predicted genes which have no similarity to any mammalian peptide, may be long non-coding RNA genes, having canonical intron-exon structure, polyadenylation sites and other coding gene-like features, but having a degenerate, or vestigial open reading frame sufficient to avoid nonsense mediated decay (Chang et al., 2007).

Of the predicted peptides, 58% were considered full length, compared to the length of the human best matching peptide, though more transcriptome sequencing is required to improve exon connectivity (**Figure 1**). Additionally, using OrthoFinder (Emms and Kelly, 2015), which identifies both "orthogroups" (genes descended from a single gene in the last common ancestor of a group of species, allowing many-to-many relationships and gene expansions) and orthologs between each pair of species in the comparison, we identified 21,136 orthogroups between *VicPac3.1*, dromedary, wild and captive Bactrian camels, cow and sheep with a mean size of 9 genes per orthogroup (**Supplementary Table 4**). Of these, 17,916 orthogroups contained an alpaca (*VicPac3.1*) protein, which provides further evidence for the quality of these gene annotations. Notably, 15,777 orthogroups contained all six species and 3,959 orthogroups contained all six species and were comprised of single-copy genes (**Supplementary Table 4**). The latter is close to the 4,104 mammalian BUSCOs[9]. The quality of the assembly and annotation was further validated by aligning all alpaca orthologs from chromosomally anchored scaffolds to the dromedary, Bactrian camel, cattle and human genomes (**Supplementary Tables 5, 6**) and compiling conserved synteny data between alpaca-human and alpaca-cattle with 11,765 and 8,494 orthologs, respectively.

Repeat content in *VicPac3.1* was 33.5% and largely the same as in *Vipacos_V1* (**Table 1**). RepeatMasker[10] based annotation (**Supplementary Table 7**) identified over 4.6 million repeat elements, with the most abundant class being LINEs (19.41%), followed by LTR elements (5.81%) and SINEs (3.79%), of which the vast majority were MIRs (3.25%). DNA transposons accounted for 3.25% of sequence and these were largely comprised of the hAT-Charlie superfamily (1.75%).

As 75.9% of scaffolds were chromosomally assigned, annotation this gave a better idea about the gene content and gene density of individual chromosomes (see **Table 3** and **Figure 2**). In total, 31,748 predicted genes were assigned to chromosomes in *VicPac3.1*. The highest number was assigned

[5]http://www.ensembl.org/index.html?redirect=no

[6]https://genome.ucsc.edu/

[7]https://www.ncbi.nlm.nih.gov/genome

[8]https://www.ncbi.nlm.nih.gov/refseq/

[9]https://busco.ezlab.org/

[10]http://www.repeatmasker.org/

TABLE 3 | Chromosomal assignment of *VicPac3.1* showing per each chromosome assembly size, number of unique assigned scaffolds, number of annotated genes, gene density, and human homology.

	VicPac3.1					*VicPac2.0.2*	
Alpaca chr.	**Assembly size, bp**	**No. of unique mapped scaffolds**	**No. of genes**	**Genes per Mb**	**Human homology**	**Assembly size, bp**	**No. of unique mapped scaffolds**
1	101,041,233	5	1625	16.1	3q, 21q	41,153,578	5
2	121,370,620	1	1650	13.6	4	36,264,523	4
3	83,363,794	3	1269	15.2	5	66,866,246	7
4	65,636,945	3	1166	17.8	9	36,674,619	6
5	96,274,254	1	1428	14.9	2q	67,623,744	5
6	74,791,714	2	1448	19.6	14q, 15q	34,188,095	5
7	31,168,711	1	641	2.1	7	9,531,993	3
8	70,270,077	1	1028	14.7	6q	62,544,616	5
9	74,791,714	3	1081	14.4	1p, 16q, 19q	26,984,682	4
10	39,582,034	1	794	19.9	11	0	0
11	77,176,758	6	1958	25.4	1q, 10q	36,454,531	6
12	48,986,614	2	910	18.6	12q	30,043,790	2
13	61,008,235	3	1491	2.4	1p	55,336,466	6
14	67,111,318	2	901	13.4	13q	19,497,535	4
15	32,418,436	2	643	2.0	2p	23.521.627	3
16	39,074,364	6	1220	3.1	10p, 17q	36,989,750	8
17	46,944,759	1	887	18.9	3p	40,598,934	2
18	29,910,177	2	930	31.0	7, 16p	24,952,824	2
19	24,022,313	1	693	28.9	20q	12,494,946	1
20	38,741,345	2	1110	28.5	6p	15,672,241	2
21	29,520,914	3	895	30.9	1q, 16q*	15,741,292	4
22	25,522,599	1	891	34.3	5q, 19p	26,415,928	2
23	29,440,657	2	520	17.9	1q, 13q	32,337,675	3
24	18,346,407	1	318	17.7	18	15,189,407	2
25	60,195,357	3	1467	24.5	8q	26,317,781	5
26	27,987,978	2	537	19.2	4q, 8p	32,483,939	2
27	22,699,463	1	11	0.5	15q	8,774,263	2
28	16,162,605	1	412	25.8	2p	13,182,695	2
29	16,162,605	2	461	28.8	8q	24,598,565	2
30	13,130,742	3	238	18.3	18q	11,278,592	3
31	13,602,737	1	323	23.1	4, 8p	12,583,175	1
32	22,732,685	3	677	29.4	12q, 22q	8,370,595	3
33	16,261,182	1	451	28.2	11q	12,417,884	2
34	22,097,801	1	526	23.9	12p	16,301,232	2
35	18,484,027	4	397	22.1	10p	10,673,185	4
36	5,377,765	3	64	12.8	7p	3,455,596	4
X	36,971,808	8	687	17.2	X	16,549,574	6

**Denotes novel conserved synteny, not detected by Balmus et al. (2007); comparative information for VicPac2.0.2 is presented in the two far right columns.*

to chr11 (1,958 genes), followed by chr2 and chr1 (**Table 3** and **Figure 2A**). Chromosomes 1 and 2 had the longest assemblies, thus it was unsurprising that they contained the most predicted genes. On average, there was a predicted gene every 103 kb of chromosomally assigned sequence, with chr22 being the most gene dense (34.3 genes/Mb) while the most gene sparse chromosome was chr27 with 0.5 genes per Mb (**Table 3** and **Figure 2B**). These numbers, however, are expected to change when the annotation improves and more of the currently unassigned scaffolds will be mapped to chromosomes.

Highlights of Selected Features of the Alpaca Genome

The Major Histocompatibility Complex (MHC)

We specifically examined the sequence of the alpaca MHC and characterized MHC organization and gene content in relation to other camelids and cetartiodactyls. The region encodes many proteins of the innate and adaptive immune systems and contains the key immune response genes for host-pathogen interactions (Trowsdale, 1995; Kelley and Trowsdale, 2005; Plasil et al., 2016; Viluma et al., 2017). In order to

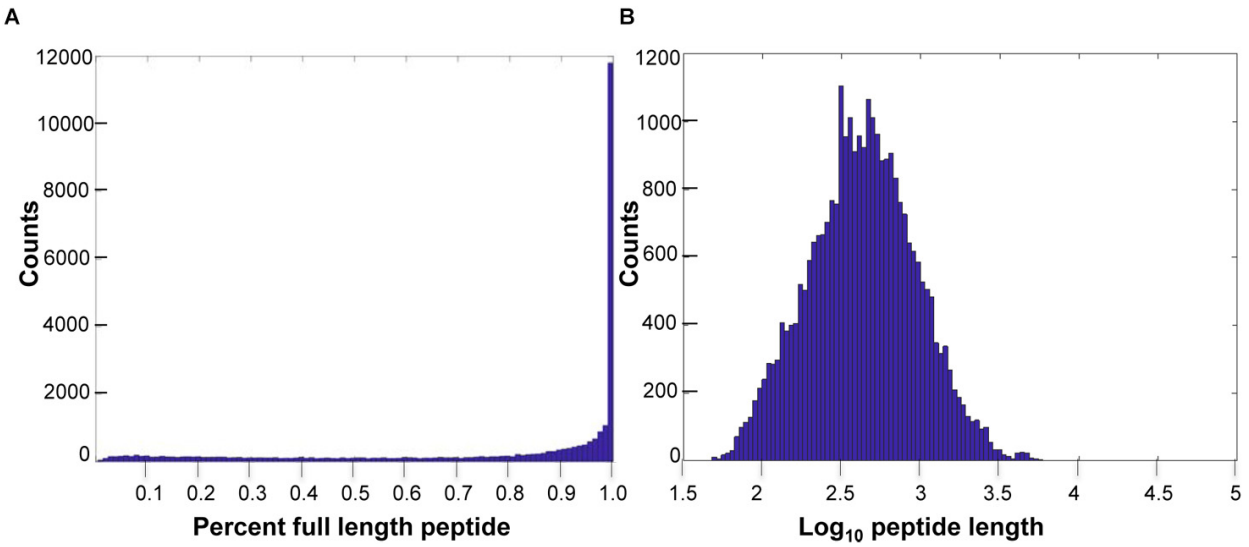


FIGURE 1 | Predicted peptide length distributions. **(A)** Histogram of *VicPac3.1* peptide annotation showing number of peptides full length, as defined by 90% or better overlap of the closest human peptide match. **(B)** Histogram of predicted peptide lengths (log10) for *VicPac3.1*.

counteract the high variability of pathogens and pathogen-derived molecules, the MHC has evolved into one of the most gene rich, highly polymorphic, and structurally complex regions of the mammalian genome, and is characterized by copy number variations and segmental duplications (Kelley and Trowsdale, 2005; Viluma et al., 2017). This complexity means that this relatively small region, which spans only approximately 4 Mb (Kelley and Trowsdale, 2005), is among the most challenging regions for genome assembly and annotation.

The MHC is located in camelid chr20, as revealed by cytogenetic mapping of specific MHC loci in alpaca (Avila et al., 2014a) and dromedary (Plasil et al., 2016). In *VicPac3.1*, two large scaffolds of 21.1 Mb (ScfyRBE_77293) and 17.6 Mb (ScfyRBE_9351) (**Table 3** and **Supplementary Table 2**) were uniquely mapped to chr20. This resulted in a 38.7 Mb assembly for chr20, making it among the largest and most contiguous assemblies of the medium size alpaca chromosomes (**Table 3**). The assembly of chr20 in *VicPac3.1*, also served as a good scaffold for placing dromedary and Bactrian camel assemblies on the chromosome, allowing for detailed comparison of the MHC region in New and Old World camelids (**Figure 3**).

The alpaca MHC spanned two separate scaffolds: Class I (723 kb) and Class III were in ScfyRBE_77293 and Class II (320 kb) was in ScfyRBE_9351 (**Figure 3**). The overall organization of the alpaca MHC relative to the centromere-telomere axis closely resembled the MHC of the dromedary and Bactrian camel (Plasil et al., 2016), i.e., centromere-Class II-Class III-Class I-telomere (**Figure 3A**). This orientation was confirmed with the cytogenetic and sequence map position of the *CRISP2* gene, which does not belong to MHC, maps very close to the centromere in chr20q (Avila et al., 2014a), and was found in scaffold ScyRBE_9351 together with MHC Class II sequences (**Figure 3A**). The MHC organization, like that seen in camelids where all MHC genes are syntenic and Class III sequences are positioned between Class I and Class II sequences, is typical of all mammalian (Ruan et al., 2016a,b; Viluma et al., 2017) and many vertebrate species (Flajnik, 2018). Alpaca and camel Class II sequences seem to be present in one block as seen in humans, pigs and horses (Li et al., 2012; Viluma et al., 2017). This is in contrast to cattle, sheep and porpoise where Class II has been disrupted by a large inversion into IIa and IIb sub-regions that happened in the ancestral chromosome of these cetartiodactyl lineages (Childers et al., 2006; Gao et al., 2010; Li et al., 2012; Ruan et al., 2016a).

Further, we more closely inspected the gene contents and order of Class I and Class II genes in alpaca, Bactrian camel and dromedary (**Figure 3**). In general, these MHC regions were collinear in camelids, though a few differences between the species were observed. In Class I, seven genes were annotated in alpaca and Bactrian camel but nine genes in the dromedary, because of an expansion of *HLA-A*-like sequences in the latter (**Figure 3C**). We speculate that the unique and specialized microbiomes of deserts (Bang et al., 2018) may have driven expansion of HLA-A in this genome. Further, no sequences of *HLA-G* corresponding to Class I heavy chain paralogs were found in alpaca, though these sequences are present in both camel species. In contrast Class I heavy chain paralogs *HLA-E* and *HLA-B* were present only in alpaca and not in camels. Class II contained 16 genes in alpaca and 17 genes in camels (**Figure 3B**). The difference was due to the *HLA-DRB1* locus, which was found in camels but not in the alpaca. Furthermore, an inversion has probably happened in the dromedary Class II changing the relative order of *HLA-DRB1* and *HLA-DRB4* in relation to that in the Bactrian camel. These minor differences in MHC gene content between camelids may be true, though it is equally plausible that they are due to difficulties associated with annotation of the highly variable MHC sequences. However, a general observation about the camelid Class II region was that even though the region is collinear between the three species,

FIGURE 2 | Features of alpaca chromosomes in *VicPac3.1*. **(A)** y-axis: the number of predicted genes per chromosome; x-axis: total length of chromosomally assigned scaffolds; **(B)** Gene density per chromosome; magenta – chromosomes with gene density < 5 genes/Mb; dark blue – chromosomes with gene density > 25 genes/Mb.

the relative positions of genes and distances between them are similar in camels but different in alpacas (**Figure 3B**), suggesting possible independent segmental duplications or expansion of gene families in the two camelid lineages.

Chromosome 36 – A Candidate Region for the Minute Chromosome Syndrome (MCS)

Chromosome 36 is the smallest autosome in the alpaca genome and of interest for several reasons. First, it was the only

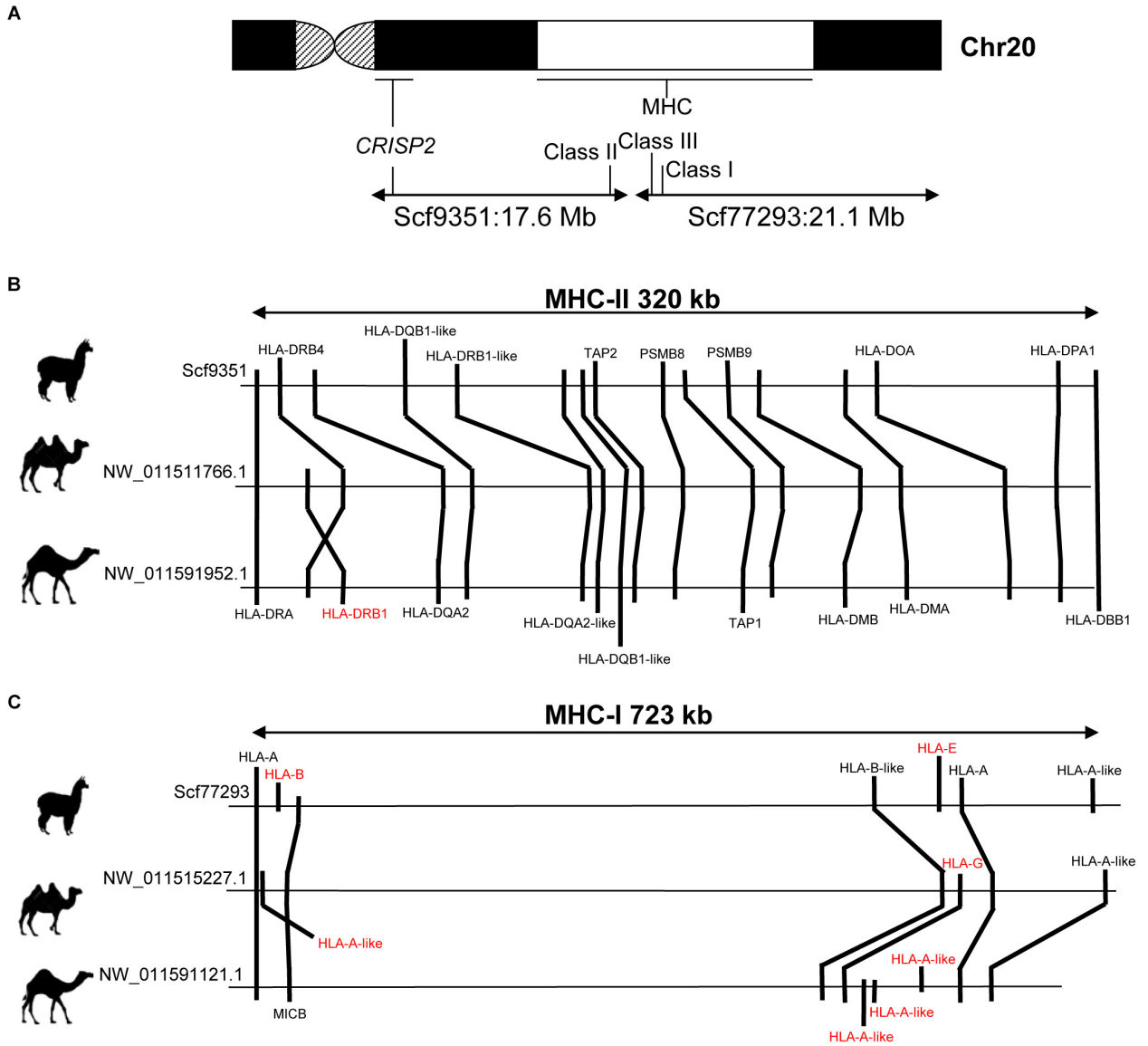


FIGURE 3 | Comparison of MHC in alpaca, dromedary and Bactrian camel. **(A)** Schematic of camelid chr20 showing FISH map positions of MHC and *CRISP2* and corresponding scaffolds in *VicPac3.1*; **(B)** comparative organization of MHC Class II; **(C)** comparative organization of MHC Class I; gene symbols in red denote loci that show differences between species; relative positions of genes and distances between loci correspond to the sequence map.

chromosome in camelid genomes for which human and other mammalian homology could not be revealed by Zoo-FISH (Balmus et al., 2007). Because of that, it was suggested that the chromosome is largely heterochromatic and contains very few genes if any at all. This hypothesis changed when two protein coding genes, *ZPBP* and *VWC2*, were mapped to chr36 by FISH revealing its homology to a segment in the short arm of human chr7 (HSA7p) at 49–56 Mb (Avila et al., 2014a, 2015). The second and perhaps even more important reason for interest in chr36 is its involvement in the *Minute Chromosome Syndrome (MCS)* (Avila et al., 2015). This is a unique chromosomal abnormality in alpacas and llamas with no counterpart in any other mammalian species. The condition is invariably associated with abnormal sexual development and infertility in females and curiously, it is recurrent (Avila et al., 2014b; Fellows et al., 2014; Raudsepp and Chowdhary, 2016). Cytogenetic manifestation of *MCS* is a dramatic size difference between the homologs of chr36, and because it was thought that the smaller homolog is abnormal, the condition was named *minute*. Now it is known that the abnormal one is the larger chr36, because it carries a massive nucleolus organizer region (NOR), which is not found in chr36 in normal alpacas (Avila et al., 2015). Molecular causes of *MCS* are, however, poorly understood and an improved reference assembly for chr36 is an important resource to study the condition at molecular level. In *VicPac3.1*, chr36 is represented by three scaffolds with a total size of about 5 Mb (**Table 4**). Altogether, we predicted

TABLE 4 | Chromosome 36 scaffolds and predicted genes with orthologs in HSA7.

Scaffold	Scaffold size, bp	Gene symbol	HSA7 sequence position
ScfyRBE_2631	352,287	*IGFBP1*	chr7:45,888,357-45,893,668
		IGFBP3	chr7:45,912,598-45,921,274
ScfyRBE_77331	2,828,351	*TNS3*	chr7:47,275,154-47,582,144
		HUS1	chr7:47,963,288-47,979,543
		SUN3	chr7:47,987,151-48,029,119
		C7orf57	chr7:48,035,520-48,061,297
		ABCA13	chr7:48,171,460-48,647,495
		VWC2	chr7:49,773,661-49,921,950
		ZPBP	chr7:49,937,441-50,093,264
		SPATA48	chr7:50,096,036-50,159,830
		IKZF1	chr7:50,304,669-50,405,101
		FIGNL1	chr7:50,444,129-50,542,535
		DDC	chr7:50,531,759-50,543,463
		GRB10	chr7:50,592,580-50,782,567
		COBL	chr7:51,016,212-51,316,799
ScfyRBE_77323	2,197,127	*VSTM2A*	chr7:54,542,325-54,571,080
		SEC61G	chr7:54,752,253-54,759,974
		LANCL2	chr7:55,365,448-55,433,742
		VOPP1	chr7:55,470,613-55,572,525
		PGAM2	chr7:44,062,727-44,065,587
		DBNL	chr7:44,044,717-44,061,716
		URGCP	chr7:43,875,913-43,906,626
		MRPS24	chr7:43,866,558-43,869,557

The order of genes in the table follows their relative order within chr36 scaffolds.

64 genes in chr36, of which 23 have known orthologs in HSA7p showing that alpaca chr36 shares homology to a ∼12 Mb region in HSA7p (**Table 4**).

Adaptations to High Altitude

Alpacas are adapted to high altitude and low oxygen environments, and therefore different evolutionary forces must have shaped their genomes as compared to dromedary and Bactrian camels, the desert species. Therefore, we specifically aimed to identify in the alpaca genome candidate genes for high altitude adaptations. We selected 20 genes for which signatures of positive selection have been reported in other high altitude species (**Table 5**). Through the application of d_N/d_S substitution ratio ω (see Material and Methods; **Supplementary Table 8**), we investigated whether any of these genes exhibit signals of selection in camelids. Nine high altitude adaptation genes exhibited sites that were under negative (purifying) selection in the alpaca compared to other camelids (**Table 5**), suggesting selection to remove deleterious mutations that might alter gene function. Three genes in this group, *EPAS1*, *EGLN1,* and *PPARA* regulate or are regulated by hypoxia inducible factor 1α (Hif-1α), which is a master regulator of the cellular response to hypoxia (Qiu et al., 2012; Simonson et al., 2012). All three genes are known to be involved in high altitude adaptations in dogs (Gou et al., 2014), humans (Beall, 2014; Bigham and Lee, 2014; Jeong et al., 2014), and *EPAS1* also in Tibetan snakes (Li et al., 2018). *EPAS1* genotypes have been associated in several studies with the dampened hemoglobin phenotype, while noncoding variants in and around *EPAS1* and *EGLN1*, are strongly associated with a reduced blood concentration of hemoglobin in Tibetan highlanders (Beall, 2014; Bigham and Lee, 2014; Gou et al., 2014). Under purifying selection in alpacas are also genes encoding for ARG2 and ADAM17 proteins, which both affect Hif-1α stability and activity (Qiu et al., 2012). Alleles of human *ADAM17* are present at significantly different frequencies in Tibetans compared to low-altitude dwellers (Simonson et al., 2012). The *NFE2L2* gene has unique amino acid residue changes in the dromedary and Bactrian camel genomes and is correlated with the oxidative stress response (Wu et al., 2014). In the present analysis, this gene exhibits signatures of purifying selection in alpaca, but not in camels (**Table 5** and **Supplementary Table 8**). Among the candidate genes tested, only *PPARA* and *EHHADH* were under positive selection in alpacas but not in camels, showing significant higher branch specific ω value (**Supplementary Table 8**). Signatures of both purifying and positive selection were found in different regions of *NOTCH4 and RBPJ*, with both genes suggested to be involved

TABLE 5 | Candidate genes for high altitude adaptations and signatures of selection in the alpaca.

Gene symbol	Signature of selection	Species where the gene is under positive selection	References
ACAA1A	–	Deer mouse	Scott et al., 2015
ADAM17	Negative	Yak	Qiu et al., 2012
ARG2	Negative	Yak	Qiu et al., 2012
ATF6	–	Pig	Jia et al., 2016
CKMT1	–	Deer mouse	Scott et al., 2015
EFEMP1	Negative	Pig	Jia et al., 2016
EGLN1	Negative	Yak, dog, human	Qiu et al., 2012; Bigham and Lee, 2014; Gou et al., 2014; Jeong et al., 2014
EHHADH	Positive	Deer mouse	Scott et al., 2015
EPAS1	Negative	Dog, human, snakes	Bigham and Lee, 2014; Gou et al., 2014; Jeong et al., 2014; Li et al., 2018
ERP44	–	Camels (oxidative stress	Wu et al., 2014
HOXB6	–	Pig	Jia et al., 2016
IKBKG	–	Pig	Jia et al., 2016
KLF6	Negative	Pig	Jia et al., 2016
MGST2	–	Camels (oxidative stress	Wu et al., 2014
MMP3	–	Yak	Qiu et al., 2012
NFE2L2	Negative	Camels (oxidative stress	Wu et al., 2014
NOTCH4	Negative and positive	Deer mouse	Scott et al., 2015
PPARA	Positive	Deer mouse, human	Simonson et al., 2012; Scott et al., 2015
RBPJ	Negative and positive	Pig	Jia et al., 2016
SF3B1	Negative	Pig	Jia et al., 2016

in the regulation of responses to hypoxia in deer mouse (Scott et al., 2015) and pig (Jia et al., 2016).

Genes Involved in Fiber Color and Quality

Alpacas produce one of the most highly prized natural fibers in the world. This fiber comes in a large range of natural colors, which is a significant point of differentiation with fine fiber sheep, such as Merinos. The key mammalian color genes *MC1R* (melanocortin 1 receptor) and *ASIP* (agouti signaling protein) have also been found to regulate alpaca fiber color (Feeley and Munyard, 2009; Feeley et al., 2011). Interestingly, although the donor of the DNA used for the alpaca genome (*Carlotta*; **Table 1**) was fawn, the *ASIP* gene in all *VicPac* genomes, including *VicPac3.1*, contains a 57 bp deletion in exon 4 associated with loss of function of ASIP, and black color. However, this is counteracted by epistatic interaction of *MC1R*, which is homozygous for the alternative allele at two of the three known loss of function SNPs (Feeley and Munyard, 2009), and which prevents the expression of eumelanin (black pigment). Importantly, *MC1R* is correctly annotated in *VicPac3.1* vs. *VicPac2*, in which it was misnamed and annotated as having three exons instead of one. It was recently shown that alpaca and camel *MC1R* maps by FISH to chr21 (Alshanbari et al., 2019) and not to chr9 as anticipated by Zoo-FISH (Balmus et al., 2007) and FISH mapping orthologs from HSA16q (Avila et al., 2014a). The location of *MC1R* in alpaca chr21 is supported by *VicPac3.1*, showing that chr21 shares conserved synteny with both HSA1 and the terminal part of HSA16q (**Table 6**). The annotation of other important mammalian color genes such as Tyrosinase related protein 1 (*TYRP1*), dopachrome tautomerase (*DCT*), Premelanosomal protein (*PMEL*), KIT oncogene (*KIT*), KIT oncogene ligand (*KITLG*), and Solute carrier 36 A1 (*SLC36A1*), is also improved in *VicPac3.1* as compared to *VicPac2* (**Table 7**).

The new assembly also improved sequence quality, annotation and chromosomal assignment of keratin (*KRT*) and keratin-associated protein (*KRTAP*) genes, some of which are the primary candidates for fleece and fiber quality (Allain and Renieri, 2010). Like in other mammals, alpaca *KRT* and *KRTAP* genes are clustered in gene families and located predominantly in chr12 (25 *KRT* genes) and chr16 (22 *KRT* genes and 2 *KRTAP*s) (**Table 8**).

TABLE 6 | Alpaca chr21 scaffolds and predicted genes (*MC1R* is highlighted) with known human orthologs.

Scaffold	Gene symbol	Human location; chr: sequence position
ScfyRBE_283	*NOS1AP*	chr1:162,069,774-162,368,451
	DDR2	chr1:162,632,465-162,787,400
	TOR3A	chr1:179,081,377-179,095,996
ScfyRBE_77299	*XPR1*	chr1:180,632,004-180,890,251
	LAMC2	chr1:183,186,288-183,244,900
	EDEM3	chr1:184,690,231-184,754,913
ScfyRBE_77374	*PIEZO1*	chr16:88,715,338-88,785,220
	ZFPM1	chr16:88,453,317-88,537,016
	BANP	chr16:87,951,434-88,077,318
	IRF8	chr16:85,898,803-85,922,609
	GSE1	chr16:85,613,216-85,676,204
	FMO5	chr1:147,186,259-147,225,638
	CTSK	chr1:150,796,208-150,808,323
	NIT1	chr1:161,118,101-161,121,067
ScfyRBE_14	*GAS8*	chr16:90,022,600-90,044,971
	MC1R	chr16:89,914,847-89,920,951
	SPG7	chr16:89,490,917-89,557,748
	ANKRD11	chr16:89,285,175-89,490,318
	SLC22A31	chr16:89,195,761-89,201,664

The order of genes in the table follows their relative order in corresponding scaffolds.

Summary

Reference assembly *VicPac3.1* with its improved accuracy, contiguity, chromosomal anchoring and preliminary annotation, constitutes a key resource for understanding the architecture of the alpaca genome, and is critical for the discovery of genetic blueprints of diseases/disease resistance, congenital disorders and traits of biological significance. It will provide a strong basis for whole genome population-scale studies in alpacas and other South American camelids, and for comparative genomics among camelids and with other mammals. High quality assembly is also the prerequisite for in depth functional annotation of the alpaca genome in the future, similar to the FAANG initiatives that are ongoing in other domestic species (Andersson et al., 2015).

MATERIALS AND METHODS

Ethics Statement

Procurement of blood and tissue samples followed the United States Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training. These protocols were approved as AUP #2011-96, # 2018-0342 CA and CRRC #09-47 at Texas A&M University.

Samples and DNA Isolation

Peripheral blood was procured from a female Huacaya alpaca *Nyala's Accoyo Empress Carlotta* – the same DNA donor that was used for the assemblies VicPac1 and VicPac2. Blood DNA was isolated using a Gentra Puregene Blood Kit (Qiagen), following the manufacturer's protocol, and evaluated for quality and quantity on the Agilent 2200 TapeStation. This showed that the gDNA was of high quality and high molecular weight (i.e., fragments larger than 50,000 bp) and suitable not only for Illumina sequencing, but also for both long-read Pacific Biosciences (PacBio) sequencing and constructing Chicago® libraries for HiRise™ scaffolding (Dovetail Genomics).

Genome Sequencing and Assembly

Two sequencing libraries were generated: a shotgun paired – end library (fragment size 200 bp) and a mate-pair library (2–5 kb) which were sequenced across 8 lanes on the Illumina 2500 platform (5 lanes paired-end and 3 lanes mate-pair; both 2 × 100 bp) commercially at Macrogen Inc. (SK) to generate short-read sequence data with > 100X genome coverage. Additionally, one PacBio SMRT cell library was

TABLE 7 | Select mammalian coat color genes in *VicPac3.1*.

Gene symbol	*VicPac3.1* no. of exons	*VicPac2.0* no. of exons	Scaffold *VicPac3.1*	Alpaca chr. *VicPac3.1*	FISH; Avila et al., 2014a
MC1R	1	3	ScfyRBE_14	21	n/a
TYRP1	8	7	ScfyRBE_2524	4	4q21-q22
DCT	8	7	ScfyRBE_4179	14	n/a
PMEL	12	7	ScfyRBE_77320	16	n/a
KIT	22	19	ScfyRBE_26	2	2q24
KITLG	10	11	ScfyRBE_77306	12	12q22
SLC36A1	14	10	ScfyRBE_5827	3	3q12

constructed and sequenced across 20 SMRT cells on the RSII platform to generate long-read data with 5-6X genome coverage; conducted commercially by PacBio Sequencing Services at the University of Washington. Short reads were filtered for quality, adaptors removed and filtered for a minimum length of 60 bp using Trimmomatic v0.33 (Bolger et al., 2014) with ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 SLIDINGWINDOW:4:25 MINLEN:60. The final genome assembly was produced through a multi-stage process. First, we generated a hybrid assembly with MaSuRCA v3.2.1 (Zimin et al., 2013) using default parameters, using all the paired-end and mate-pair Illumina data, ∼22X Roche 454 data from *VicPac2.0* and ∼3X Sanger data from *VicPac1* (both available via PRJNA30567). The hybrid assembly was designated as *Qmas1*. This was further developed into a meta-assembly guided by the *VicPac2* assembly and Illumina mate-pair alignments using Metassembler (Wences and Schatz, 2015) with MUMmer4 (Marcais et al., 2018) and the following parameters: *MateAn_A = 2000, MateAn_B = 3000, nucmer_l = 50, nucmer_c = 300*) in order to generate a more contiguous assembly for the alpaca genome. This assembly was designated as *Qmas1/VicPac2*. Using the ∼5X PacBio consensus sub-reads and Illumina mate-pair data, we scaffolded the *Qmas1/VicPac2* contigs using OPERA-LG v.2.05 (Gao et al., 2016); assembly designation *VicPac3.0*. The final assembly, *VicPac3.1*, was obtained by constructing two 2X 151 bp Chicago® libraries which were then subjected to HiRise™ scaffolding along with *VicPac3.0* and the mate-pair libraries, and this was done by Dovetail Genomics (United States).

TABLE 8 | Clusters of keratin and keratin-associated protein genes in *VicPac3.1*.

Gene symbol	*VicPac3.1* Scaffold	Chromosome
KRT18; KRT8; KRT78; KRT79; KRT4; KRT3; KRT77; KRT1; KRT2; KRT73; KRT72; KRT74; KRT71; KRT5; KRT6A; KRT75; KRT82; KRT84; KRT85; KRT83; KRT86; KRT81; KRT82; KRT7; KRT80	ScfyRBE_77306	12
KRT17; KRT16; KRT14; KRT9; KRT19; KRT13/15; KRT36; KRT35; KRT32; KRT31; KRT33A; KRTAP3-1; KRTAP3-3; KRT40; KRT39; KRT23; KRT20; KRT12; KRT27; KRT26; KRT25; KRT24; KRT222;	ScfyRBE_77388	16
KRT6C; KRT6B	ScfyRBE_2857	n/a
KRTAP13-1; KRTAP13-2; KRTAP7-1	ScfyRBE_4	1

Genes are ordered in the table following their relative order in corresponding scaffolds.

Chromosomal Assignment

Sequence scaffolds were anchored to alpaca autosomes and the X chromosome with the help of alpaca cytogenetic markers (Avila et al., 2014a,b, 2015). Sequences of the overgo and PCR primers that were used for FISH analysis were mapped to *VicPac3.1* scaffolds using BBMap v35[11] with the following parameters: *pairedonly = t, minid = 0.97 and pairlen = 500*, which generated no ambiguous mappings, retaining those with 97% identity (equivalent of 1 bp mismatch) and that the primers were mapped in the correct orientation. Overgo primers were mapped with the same parameters, but allowing for 95% identity matches. We considered scaffolds anchored when the primers uniquely mapped to one scaffold and the primers mapped in the correct orientation.

Preliminary Genome Annotation

A preliminary annotation of the *VicPac3.1* was produced primarily for comparative assessment. We used AUGUSTUS v3.3.1 (Hoff and Stanke, 2018) for gene model prediction. First, gene hints were built using VicPac2.0.2 (GCA_000164845.3), human (GCA_000001405.27), cow (*Bos taurus*; GCA_000003055.3), dromedary (*Camelus dromedaries*; GCA_000767585.1), and Bactrian camel (GCF_000311805.1 and GCF_000767855.1) peptides. Peptides were mapped to the alpaca genome assembly using BLAT v. 36x2 (Kent, 2002) with default parameters, then converted to hints with blat2hints.pl, taking the best two matches to every peptide. AUGUSTUS was then run with these hints, and human was used as the training species (closest pre-trained species). Finally, the predicted gene models were confirmed by mapping testis and skin RNA-Seq data to *VicPac3.1* with STAR v 2.5 (Dobin et al., 2013) in two-pass mode with default parameters and checking correct junctions in the STAR junction files and well-known gene models in IGV (Robinson et al., 2011; Thorvaldsdottir et al., 2013).

Interspersed repeats were identified through a homology-based approach using RepeatMasker v4.07[12] with RMBlast v 2.2.27+[13] and TRF v4.09 (Benson, 1999) searches against Dfam

[11] http://bib.irb.hr/datoteka/773708.Josip_Maric_diplomski.pdf

[12] http://www.repeatmasker.org/

[13] http://www.repeatmasker.org/RMBlast.html

2.0, Dfam consensus (Hubley et al., 2016) and Repbase (Bao et al., 2015) databases, both 20170127 issue.

Genome Contiguity and Completeness Assessment

The contiguity and completeness of the alpaca genome assemblies were evaluated using several methods. We computed core assembly metrics (N50, L50, number of scaffold, longest scaffold, GC content and proportion N's) for our 4 assemblies, *VicPac2* and the alpaca assembly (Vi_pacos_V1) (Wu et al., 2014). Completeness of the four assemblies generated in this study was directly compared using BUSCO (Simao et al., 2015) analysis of conserved orthologs. BUSCO score comparisons between organisms can serve as useful benchmarks for assembly completeness so we also compiled BUSCO assessments for cow (v3.1.1; GCF_000003055.6), sheep (*Ovis aries*; v4.0; GCF_000298735.2) dromedary (GCF000767585.1), and Bactrian (both domesticated, GCF_00767855.1, and wild, GCF_000311805.1) camels. We ran BUSCO v3.0.2[14] with *geno* mode, mammalia_odb9, Blast v2.2.26+ (Camacho et al., 2009), HMMer v3.1 (Eddy, 2011) and Augustus v3.2. Lastly, we compared gene model predictions at the protein level for *VicPac3.1*, *VicPac2*, and Vi_pacos_V1 using both standard and reciprocal best-hit blastp, using Blast v2.2.26+ and an evaluated cut off of e^{-10}.

Comparative Analysis

We used OrthoFinder v1.1.10 (Emms and Kelly, 2015) with Diamond v0.9.9.110 (Buchfink et al., 2015), and FastME v2.1.5 (Lefort et al., 2015) to identify orthologs, orthogroups, paralogs and compute gene (ortholog) and species trees in an all vs. all comparison of *VicPac3.1*, dromedary, and both wild and domesticated Bactrian camels, cow and sheep.

The *VicPac3.1* scaffolds anchored to chr20 were used to anchor dromedary and Bactrian scaffolds to chr20, using reciprocal best-hit Blast v2.2.26+, blastn implemented with default settings. Pairwise comparative alignments were conducted for anchored chr20 scaffolds using MAUVE v 2.4.0 (Darling et al., 2004) with default settings for alpaca (*VicPac3.1*) vs. dromedary, alpaca vs. Bactrian, alpaca vs. cow and alpaca vs. sheep. Cow and sheep genome assemblies are already chromosomally assigned so we used their respective chr20 sequence fasta. We compared Major Histocompatibility complex (MHC) gene synteny among camelid chr20 (alpaca, dromedary and Bactrian camels), using orthology and syntenic position between anchoring orthologs. All MHC and MHC-like genes (any gene with a blast *e*-value less than 1e-20 to any human MHC gene) in the MHC Class I and MHC Class II syntenic regions were annotated with respect to human peptide best matches.

Positive Selection

To investigate whether candidate genes involved in adaptation to high altitudes exhibit signals of selection among the Camelidae, coding sequences (CDS) of 23 genes previously identified as potentially having a role in high altitude adaptation (*EGLN1, EPAS1, PPARA, IKBKG, KLF6, RBPJ, SF3B1, EFEMP1, HOXB6, ATF6, ADAM17, MMP3, ARG2, ERP44, NFE2L2, MGST2, AQP1, AQP2, AQP3, CKMT1, EHHADH, ACAALA, NOTCH4)* were extracted from the *VicPac3.1* and from dromedary and wild and domesticated Bactrian camel assemblies. Multiple sequence alignments were conducted using GUIDANCE2 (Sela et al., 2015) with default quality cutoffs, codon alignments with PRANK (Loytynoja and Goldman, 2010) as the MSA program specified with the –F parameter. We used the longest CDS of a gene for alignment when there was more than one per species. Signatures of selection were searched with two d_N/d_S based tests using HyPhy[15] 3pc (Pond et al., 2005). First, the aBSREL (Smith et al., 2015) branch-site model, which tests if each branch in the phylogeny has a proportion of sites evolving under positive selection, as we tested all branches we FDR corrected the likelihood-ratio test *p*-values. Second, the FEL (Kosakovsky Pond and Frost, 2005) model which assumes selective pressure is constant for each site across the phylogeny and calculates whether the nonsynonymous (d_N) substitution rate is significantly different from the synonymous (d_S) rate, using the likelihood ratio test.

Transcriptome Sequencing and Analysis

High quality (RIN > 9.6) RNA was extracted from the testis of one normal male alpaca and one normal male llama using PureLink RNA Mini Kit (Ambion). The RNA was converted into cDNA with NEXTflex Rapid Directional qRNA-Seq kit (BIOO), prepared into 2 × 100 bp PE TruSeq libraries (Illumina), and sequenced on Illumina HiSeq2500 platform. We obtained, on average, 90 million PE reads per sample. The RNA from the skin samples was prepared as reported in Cransberg Ph.D. Thesis (Cransberg, 2017). Briefly, skin biopsies were collected from 20 white, 20 brown and 5 black alpacas, the RNA was extracted using Trizol reagent and the FastPrep system (Thermo Life Sciences) and an RNeasy Kit (Qiagen). After confirmation of RNA quality (Bioanalyser; Agilent) three equi-molar pools of RNA were prepared (one for each color). Sequencing libraries were prepared using an Illumina Tru-seq RNA kit, and sequenced on a single lane of an Illumina Genome Analyser GAIIx to generate 54 bp PE reads.

Genome Size Estimation

Genome size was first estimated using filtered short-read *k*-mer distributions. *k*-mer frequencies were calculated using Jellyfish v2.2.8 (Marcais and Kingsford, 2011) with canonical *k*-mers, for a range of *k*-values (17, 21, 25, and 31). These *k*-mer distributions were then analyzed in Genoscope[16] (Vurture et al., 2017) with a maximum *k*-mer coverage of 1,000 and –1 (where –1 is no maximum coverage). Genome size was also estimated by flow cytometry using the protocol described elsewhere (Zhu et al., 2012) with a modification that the concentration of RNAse was doubled (200 μg/ml). Briefly, primary fibroblast cell lines of 2 alpacas and 2 horses were cultured in T25 culture flasks until 100% confluency. The cells of each individual were trypsinized, washed 6 times in PBS, fixed in cold 70% ethanol and stained

[14] https://vcru.wisc.edu/simonlab/bioinformatics/programs/busco/BUSCO_v3_userguide.pdf

[15] http://www.hyphy.org

[16] http://qb.cshl.edu/genomescope/

with Propidium Iodine. The stained cells were analyzed on a BD Accuri^{TM} C6 personal flow cytometer separately for each animal. Results were gated in order to prevent exogenous DNA from lysed cells from affecting the results. Peaks were observed based on the amount of PI absorbed by each cell population (**Supplementary Figure 2**). Both horses and one alpaca were measured on 3 separate occasions, the other alpaca was only measured once due to a limited number of cells. The average median PI concentration and a 95% CI was calculated for the horse an alpaca, respectively, using all measurements available. The genome size was estimated using the formula: $Size_{alpaca} = PI_{alpaca}/PI_{horse}$ (2.7 Gb). Where PI_{alpaca} denotes the median amount of PI absorbed by alpaca cells; PI_{horse} denotes the median amount of PI absorbed by horse cells; 2.7 Gb is the expected size of the horse genome (Wade et al., 2009).

ETHICS STATEMENT

Procurement of blood and tissue samples followed the United States Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research and Training. These protocols were approved as AUP #2011-96, # 2018-0342 CA and CRRC #09-47 at Texas A&M University.

AUTHOR CONTRIBUTIONS

TR, BA, and PP designed and initiated the project. MR, KM, LC, FJ, TA, FA, MJ, GW, RC, and TR conducted the experimental work. MR, KM, FJ, LC, and TA carried out the genome assembly, annotation, and bioinformatics analyses. MJ and FA contributed to the testis transcriptome and data analyses. TR, AT, RC, and KM collected the samples for genome and transcriptome sequencing. TR, MR, KM, and LC wrote the manuscript with input from all authors.

FUNDING

This study was supported by grants from the Morris Animal Foundation (D09LA-004, D14LA-005) and the Alpaca Research Foundation; funds from the Curtin University, School of Biomedical Sciences supported PacBio and Dovetail sequencing and skin transcriptome analysis; the authors highly appreciate donations to ARF and MAF by Leslie Herzog of Herzog Alpacas.

ACKNOWLEDGMENTS

We thank Dr. Andrew Merriwether for providing the blood sample from *Nyala's Accoyo Empress Carlotta.*

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00586/full#supplementary-material

REFERENCES

Allain, D., and Renieri, C. (2010). Genetics of fibre production and fleece characteristics in small ruminants, angora rabbit and south american camelids. *Animal* 4, 1472–1481. doi: 10.1017/s1751731110000029

Alshanbari, F., Castaneda, C., Juras, R., Hillhouse, A., Mendoza, M. N., Gutiérrez, G. A., et al. (2019). Comparative FISH-mapping of *MC1R*, *ASIP* and *TYRP1* in New and Old World camelids and association analysis with coat color phenotypes in the dromedary (*Camelus dromedarius*). *Front. Genet.* 16:340. doi: 10.3389/fgene.2019.00340

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., et al. (2015). Coordinated international action to accelerate genome-to-phenome with faang, the functional annotation of animal genomes project. *Genome Biol.* 16:57.

Avila, F., Baily, M. P., Merriwether, D. A., Trifonov, V. A., Rubes, J., Kutzler, M. A., et al. (2015). A cytogenetic and comparative map of camelid chromosome 36 and the minute in alpacas. *Chromosome Res.* 23, 237–251. doi: 10.1007/s10577-014-9463-3

Avila, F., Baily, M. P., Perelman, P., Das, P. J., Pontius, J., Chowdhary, R., et al. (2014a). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*lama pacos*). *Cytogenet. Genome. Res.* 144, 196–207. doi: 10.1159/000370329

Avila, F., Das, P. J., Kutzler, M., Owens, E., Perelman, P., Rubes, J., et al. (2014b). Development and application of camelid molecular cytogenetic tools. *J. Hered.* 105, 858–869.

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C., Alkalaeva, E. S., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative cetartiodactyla ancestral karyotype. *Chromosome Res.* 15, 499–515.

Fellows, E., Kutzler, M., Avila, F., Das, P. J., and Raudsepp, T. (2014). Ovarian dysgenesis in an alpaca with a minute chromosome 36. *J. Hered.* 105, 870–874. doi: 10.1093/jhered/ess069

Bang, C., Dagan, T., Deines, P., Dubilier, N., Duschl, W. J., Fraune, S., et al. (2018). Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? *Zoology (Jena)* 127, 1–19. doi: 10.1016/j.zool.2018.02.004

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11.

Barreta, J., Gutierrez-Gil, B., Iniguez, V., Saavedra, V., Chiri, R., Latorre, E., et al. (2013). Analysis of mitochondrial DNA in bolivian llama, alpaca and vicuna populations: a contribution to the phylogeny of the south american camelids. *Anim. Genet.* 44, 158–168. doi: 10.1111/j.1365-2052.2012.02376.x

Beall, C. M. (2014). Adaptation to high altitude: phenotypes and genotypes. *Ann. Rev. Anthropol.* 43, 251–272. doi: 10.1146/annurev-anthro-102313-030000

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bianchi, N. O., Larramendy, M. L., Bianchi, M. S., and Cortes, L. (1986). Karyological conservation in south american camelids. *Experientia* 42, 622–624. doi: 10.1007/bf01955563

Bigham, A. W., and Lee, F. S. (2014). Human high-altitude adaptation: forward genetics meets the hif pathway. *Genes Dev.* 28, 2189–2204. doi: 10.1101/gad.250167.114

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bruford, M. W., Bradley, D. G., and Luikart, G. (2003). DNA markers reveal the complexity of livestock domestication. *Nat. Rev. Genet.* 4, 900–910. doi: 10.1038/nrg1203

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Chang, Y. F., Imam, J. S., and Wilkinson, M. F. (2007). The nonsense-mediated decay rna surveillance pathway. *Annu. Rev. Biochem.* 76, 51–74. doi: 10.1146/annurev.biochem.76.050106.093909

Childers, C. P., Newkirk, H. L., Honeycutt, D. A., Ramlachan, N., Muzney, D. M., Sodergren, E., et al. (2006). Comparative analysis of the bovine mhc class iib sequence identifies inversion breakpoints and three unexpected genes. *Anim. Genet.* 37, 121–129. doi: 10.1111/j.1365-2052.2005.01395.x

Cohen, J. (2018). Llama antibodies inspire gene spray to prevent all flus. *Science* 362:511. doi: 10.1126/science.362.6414.511

Cransberg, R. (2017). *Insights Into the Alpaca Skin Transcriptome in Relation to Fibre Colour: School of Biomedical Science*. Perth, WA: Curtin University.

Cruz, A., Cervantes, I., Burgos, A., Morante, R., and Gutierrez, J. P. (2017). Genetic parameters estimation for preweaning traits and their relationship with reproductive, productive and morphological traits in alpaca. *Animal* 11, 746–754. doi: 10.1017/s175173111600210x

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Emms, D. M., and Kelly, S. (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.

Feeley, N. L., Bottomley, S., and Munyard, K. A. (2011). Three novel mutations in asip associated with black fibre in alpacas (*Vicugna pacos*). *J. Agric. Sci.* 149, 529–538. doi: 10.1017/s0021859610001231

Feeley, N. L., and Munyard, K. A. (2009). Characterisation of the melanocortin-1 receptor gene in alpaca and identification of possible markers associated with phenotypic variations in colour. *Anim. Product. Sci.* 49, 675–681.

Flajnik, M. F. (2018). A cold-blooded view of adaptive immunity. *Nat. Rev. Immunol.* 18, 438–453. doi: 10.1038/s41577-018-0003-9

Flajnik, M. F., Deschacht, N., and Muyldermans, S. (2011). A case of convergence: why did a simple alternative to canonical antibodies arise in sharks and camels? *PLoS Biol.* 9:e1001120. doi: 10.1371/journal.pbio.1001120

Gao, J., Liu, K., Liu, H., Blair, H. T., Li, G., Chen, C., et al. (2010). A complete DNA sequence map of the ovine major histocompatibility complex. *BMC Genomics* 11:466. doi: 10.1186/1471-2164-11-466

Gao, S., Bertrand, D., Chia, B. K., and Nagarajan, N. (2016). Opera-lg: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. *Genome Biol.* 17:102.

Genome 10K Community of Scientists. (2009). Genome 10k: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* 100, 659–674. doi: 10.1093/jhered/esp086

Gou, X., Wang, Z., Li, N., Qiu, F., Xu, Z., Yan, D., et al. (2014). Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res.* 24, 1308–1315. doi: 10.1101/gr.171876.113

Griffin, L. M., Snowden, J. R., Lawson, A. D., Wernery, U., Kinne, J., and Baker, T. S. (2014). Analysis of heavy and light chain sequences of conventional camelid antibodies from *Camelus dromedarius* and camelus bactrianus species. *J. Immunol. Methods* 405, 35–46. doi: 10.1016/j.jim.2014.01.003

Hoff, K. J., and Stanke, M. (2018). Predicting genes in single genomes with augustus. *Curr. Protoc. Bioinformatics* 65:e57. doi: 10.1002/cpbi.57

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., et al. (2016). The dfam database of repetitive DNA families. *Nucleic Acids Res.* 44, D81–D89.

Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D. B., Pritchard, J. K., et al. (2014). Admixture facilitates genetic adaptations to high altitude in tibet. *Nat. Commun.* 5:3281.

Jia, C., Kong, X., Koltes, J. E., Gou, X., Yang, S., Yan, D., et al. (2016). Gene co-expression network analysis unraveling transcriptional regulation of high-altitude adaptation of tibetan pig. *PLoS One* 11:e0168161. doi: 10.1371/journal.pone.0168161

Kelley, J., and Trowsdale, J. (2005). Features of mhc and nk gene clusters. *Transpl. Immunol.* 14, 129–134. doi: 10.1016/j.trim.2005.03.001

Kent, W. J. (2002). Blat–the blast-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kosakovsky Pond, S. L., and Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22, 1208–1222. doi: 10.1093/molbev/msi105

Lefort, V., Desper, R., and Gascuel, O. (2015). Fastme 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32, 2798–2800. doi: 10.1093/molbev/msv150

Li, G., Liu, K., Jiao, S., Liu, H., Blair, H. T., Zhang, P., et al. (2012). A physical map of a bac clone contig covering the entire autosome insertion between ovine mhc class iia and iib. *BMC Genomics* 13:398. doi: 10.1186/1471-2164-13-398

Li, J. T., Gao, Y. D., Xie, L., Deng, C., Shi, P., Guan, M. L., et al. (2018). Comparative genomic investigation of high-elevation adaptation in ectothermic snakes. *Proc. Natl. Acad. Sci. U.S.A.* 115, 8406–8411. doi: 10.1073/pnas.1805348115

Loytynoja, A., and Goldman, N. (2010). Webprank: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* 11:579. doi: 10.1186/1471-2105-11-579

Marcais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). Mummer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14:e1005944. doi: 10.1371/journal.pcbi.1005944

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Marin, J. C., Rivera, R., Varas, V., Cortes, J., Agapito, A., Chero, A., et al. (2018). Genetic variation in coat colour genes mc1r and asip provides insights into domestication and management of south american camelids. *Front. Genet.* 9:487. doi: 10.3389/fgene.2018.00487

Murphy, W. J., Larkin, D. M., Everts-van, der Wind, A., Bourque, G., Tesler, G., et al. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309, 613–617. doi: 10.1126/science.1111387

Pennisi, E. (2012). Genomics. Encode project writes eulogy for junk DNA. *Science* 337:1161.

Plasil, M., Mohandesan, E., Fitak, R. R., Musilova, P., Kubickova, S., Burger, P. A., et al. (2016). The major histocompatibility complex in old world camelids and low polymorphism of its class ii genes. *BMC Genomics* 17:167. doi: 10.1186/s12864-016-2500-1

Pond, S. L., Frost, S. D., and Muse, S. V. (2005). Hyphy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679. doi: 10.1093/bioinformatics/bti079

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., et al. (2014). Refseq: an update on mammalian reference sequences. *Nucleic Acids Res.* 42, D756–D763.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., et al. (2012). The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44, 946–949.

Raudsepp, T., and Chowdhary, B. P. (2016). Chromosome aberrations and fertility disorders in domestic animals. *Annu. Rev. Anim. Biosci.* 4, 15–43. doi: 10.1146/annurev-animal-021815-111239

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Ruan, R., Ruan, J., Wan, X. L., Zheng, Y., Chen, M. M., Zheng, J. S., et al. (2016a). Organization and characteristics of the major histocompatibility complex class ii region in the yangtze finless porpoise (neophocaena asiaeorientalis asiaeorientalis). *Sci. Rep.* 6:22471.

Ruan, R., Wan, X. L., Zheng, Y., Zheng, J. S., and Wang, D. (2016b). Assembly and characterization of the mhc class i region of the yangtze finless porpoise (neophocaena asiaeorientalis asiaeorientalis). *Immunogenetics* 68, 77–82. doi: 10.1007/s00251-015-0885-7

Scott, G. R., Elogio, T. S., Lui, M. A., Storz, J. F., and Cheviron, Z. A. (2015). Adaptive modifications of muscle phenotype in high-altitude deer mice are associated with evolved changes in gene regulation. *Mol. Biol. Evol.* 32, 1962–1976. doi: 10.1093/molbev/msv076

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). Guidance2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Simonson, T. S., McClain, D. A., Jorde, L. B., and Prchal, J. T. (2012). Genetic determinants of tibetan high-altitude adaptation. *Hum. Genet.* 131, 527–533. doi: 10.1007/s00439-011-1109-3

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353. doi: 10.1093/molbev/msv022

Steward, C. A., Parker, A. P. J., Minassian, B. A., Sisodiya, S. M., Frankish, A., and Harrow, J. (2017). Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med.* 9:49.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Trowsdale, J. (1995). "Both man & bird & beast": comparative organization of mhc genes. *Immunogenetics* 41, 1–17.

Viluma, A., Mikko, S., Hahn, D., Skow, L., Andersson, G., and Bergstrom, T. F. (2017). Genomic structure of the horse major histocompatibility complex class ii region resolved using pacbio long-read sequencing technology. *Sci. Rep.* 7:45518.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). Genomescope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326, 865–867.

Wences, A. H., and Schatz, M. C. (2015). Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol.* 16:207.

Wheeler, J. C. (1995). Evolution and present situation of the south american camelidae. *Biol. J. Linn. Soc.* 54, 271–295. doi: 10.1111/j.1095-8312.1995.tb01037.x

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5:5188.

Zhou, X., Xu, S., Yang, Y., Zhou, K., and Yang, G. (2011). Phylogenomic analyses and improved resolution of cetartiodactyla. *Mol. Phylogenet. Evol.* 61, 255–264. doi: 10.1016/j.ympev.2011.02.009

Zhu, D., Song, W., Yang, K., Cao, X., Gul, Y., and Wang, W. (2012). Flow cytometric determination of genome size for eight commercially important fish species in china. *In Vitro Cell Dev. Biol. Anim.* 48, 507–517. doi: 10.1007/s11626-012-9543-7

Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The masurca genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor is currently organizing a Research Topic with one of the authors KM, and confirms the absence of any other collaboration.

Copyright © 2019 Richardson, Munyard, Croft, Allnutt, Jackling, Alshanbari, Jevit, Wright, Cransberg, Tibary, Perelman, Appleton and Raudsepp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 July 2019
doi: 10.3389/fgene.2019.00620

Natural Killer Cell Receptor Genes in Camels: Another Mammalian Model

*Jan Futas[1,2], Jan Oppelt[2,3], April Jelinek[1], Jean P. Elbers[4], Jan Wijacki[5,6], Ales Knoll[5,6], Pamela A. Burger[4] and Petr Horin[1,2]**

[1] Department of Animal Genetics, Faculty of Veterinary Medicine, University of Veterinary and Pharmaceutical Sciences, Brno, Czechia, [2] RG Animal Immunogenomics, CEITEC-VFU, University of Veterinary and Pharmaceutical Sciences, Brno, Czechia, [3] National Centre for Biomolecular research, CEITEC-MU, Faculty of Science, Masaryk University, Brno, Czechia, [4] Research Institute for Wildlife Ecology, Department of Integrative Biology and Evolution, Vetmeduni Vienna, Vienna, Austria, [5] Department of Animal Morphology, Physiology and Genetics, Faculty of Agronomy, Mendel University in Brno, Brno, Czechia, [6] RG Animal Immunogenomics, CEITEC-MENDELU, Mendel University in Brno, Brno, Czechia

OPEN ACCESS

Edited by:
Edward Hollox,
University of Leicester,
United Kingdom

Reviewed by:
Paul J. Norman,
University of Colorado Denver,
United States
Lutz Walter,
Deutsches Primatenzentrum,
Germany

***Correspondence:**
Petr Horin
horin@dior.ics.muni.cz

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 30 January 2019
Accepted: 13 June 2019
Published: 02 July 2019

Citation:
Futas J, Oppelt J, Jelinek A, Elbers JP, Wijacki J, Knoll A, Burger PA and Horin P (2019) Natural Killer Cell Receptor Genes in Camels: Another Mammalian Model. Front. Genet. 10:620. doi: 10.3389/fgene.2019.00620

Due to production of special homodimeric heavy chain antibodies, somatic hypermutation of their T-cell receptor genes and unusually low diversity of their major histocompatibility complex genes, camels represent an important model for immunogenetic studies. Here, we analyzed genes encoding selected natural killer cell receptors with a special focus on genes encoding receptors for major histocompatibility complex (MHC) class I ligands in the two domestic camel species, *Camelus dromedarius* and *Camelus bactrianus.* Based on the dromedary genome assembly CamDro2, we characterized the genetic contents, organization, and variability of two complex genomic regions, the leukocyte receptor complex and the natural killer complex, along with the natural cytotoxicity receptor genes *NCR1*, *NCR2*, and *NCR3*. The genomic organization of the natural killer complex region of camels differs from cattle, the phylogenetically most closely related species. With its minimal set of *KLR* genes, it resembles this complex in the domestic pig. Similarly, the leukocyte receptor complex of camels is strikingly different from its cattle counterpart. With *KIR* pseudogenes and few *LILR* genes, it seems to be simpler than in the pig. The syntenies and protein sequences of the *NCR1*, *NCR2*, and *NCR3* genes in the dromedary suggest that they could be human orthologues. However, only *NCR1* and *NCR2* have a structure of functional genes, while *NCR3* appears to be a pseudogene. High sequence similarities between the two camel species as well as with the alpaca *Vicugna pacos* were observed. The polymorphism in all genes analyzed seems to be generally low, similar to the rest of the camel genomes. This first report on natural killer cell receptor genes in camelids adds new data to our understanding of specificities of the camel immune system and its functions, extends our genetic knowledge of the innate immune variation in dromedaries and Bactrian camels, and contributes to studies of natural killer cell receptors evolution in mammals.

Keywords: camelid, leukocyte receptor complex, natural killer complex, SNP, microsatellites

INTRODUCTION

Camels (*Camelus* spp.) represent an important genus for a number of reasons. Due to their adaptation to desert or semi-desert regions, Old World camels tolerate harsh conditions, which are inhospitable for many livestock species, including extreme temperatures and prolonged periods without access to food and water (reviewed in Jirimutu et al., 2012). As a result, they are of socioeconomic importance across the Middle East, Northern Africa, and much of Asia, where they are used for meat, milk, hides, transportation, and sport. The significance of camels as a sustainable livestock species is likely to continue, as many regions face increased temperatures and desertification as a result of climate change (Megersa et al., 2014; Watson et al., 2016). Concurrently, recent trends towards intensive production and the movement of camel production to peri-urban settings are altering the pathogen pressures to which these animals are exposed (Abdallah and Faye, 2013).

Camels are also of importance with respect to a number of specific infectious diseases. For example, dromedaries (*Camelus dromedarius*) are a natural host of Middle East respiratory syndrome coronavirus, and transmission of the virus from camels to humans has been confirmed (Gossner et al., 2014; Hemida et al., 2017). Interestingly, significant differences exist between dromedaries and Bactrian camels (*Camelus bactrianus*) with regard to their susceptibility to foot and mouth disease (FMD), one of the most costly diseases of production animals worldwide; i.e., dromedaries are not susceptible to FMD and do not transmit infection (Wernery and Kinne, 2012). Furthermore, the immune system of camels has several unusual features. Notable among these is the presence of homodimeric heavy chain antibodies (Hamers-Casterman et al., 1993), not known to occur in any other mammalian family, which have both potential and realized applications in a variety of research, diagnostic, and therapeutic settings (Muyldermans et al., 2009; De Meyer et al., 2014; Steeland et al., 2016). The persistence of uniquely organized ileal Peyer's patches into adulthood of the dromedary (Zidan and Pabst, 2008) is another example. Additionally, productively rearranged dromedary T-cell receptor delta variable (*TRDV*) (Antonacci et al., 2011) and T-cell receptor gamma variable (*TRGV*) (Vaccarelli et al., 2012) genes undergo somatic hypermutation to generate a diversified repertoire of these genes. This mechanism has not been documented for T-cell receptor genes in other mammalian species and appears to compensate the more limited repertoire of *TRDV* and *TRGV* genes found in camels relative to other Artiodactyls (Ciccarese et al., 2014). A further atypical aspect of the camelid immunogenome is the unusually low genetic diversity of the major histocompatibility complex (MHC) of the three species of Old World camels in both class I (Plasil et al., 2019) and class II genes (Plasil et al., 2016). The immunological characterization of cellular components of the camel immune system is scarce mainly due to the small number of cross-reacting monoclonal antibodies raised against leukocyte antigens of humans (Hussen et al., 2017), bovines, and/or other related species (Mossad et al., 2006) available. This is also one of the reasons why natural killer cells and their functions in camelids have not been studied so far.

Natural killer (NK) cells constitute a heterogeneous lymphocyte population (Allan et al., 2015) involved primarily in innate immune responses against intracellular pathogens and tumor cells. They also influence adaptive immune responses *via* the production of cytokines (Vivier et al., 2011) and crosstalk with dendritic cells (Hamerman et al., 2005), play a role in placentation (Parham and Moffett, 2013), and contribute to the recognition of allogeneic cells. The diversity of the NK cell receptor repertoire is essential to the performance of these multiple functions. The integration of activating and inhibitory signals originating from various surface receptors determines the activation status of an individual NK cell, providing the capacity to discriminate between self and non-self or altered self (Lanier, 1998).

Characterization of genes underlying receptors on the NK cell surface can significantly contribute to our understanding of the functional heterogeneity of NK cells. Among them, NK cell receptors (NKR) of several gene families bind polymorphic MHC class I or MHC class I-like molecules to mediate NK cell function. Due to functional relationships between MHC and NKR molecules, the underlying genes and genomic regions represent an important biological model in terms of their co-evolution in the context of pathogen pressures, disease, and survival (Guethlein et al., 2015; Carrillo-Bustamante et al., 2016). However, the current knowledge of mammalian NKR genes, in comparison with that of MHC regions, is rather fragmentary. Two major genomic complexes encoding NKR, the natural killer complex (NKC) and the leukocyte receptor complex (LRC), have been identified in mammalian genomes. Genes in the NKC represent receptors with C-type lectin-like extracellular domains; genes in the LRC code receptors with extracellular ligand-binding domains belong to the immunoglobulin superfamily (Trowsdale et al., 2001). Despite these structural differences, some receptor families of both complexes are able to fulfil the same functions in terms of MHC class I recognition, downstream signaling, and mediation of NK cell activation/inhibition. To accomplish these ends, different NKR gene families expanded and diversified in different mammalian species, representing an example of convergent evolution in mammals (Kelley et al., 2005; Guethlein et al., 2015). Two immunologically well-defined species, humans and mice, have expanded structurally unrelated receptor families: humans use the killer-cell immunoglobulin-like receptors (KIR) and leukocyte immunoglobulin-like receptors (LILR), both encoded within the LRC, whereas in mice the killer-cell lectin-like receptor (KLR) genes of one family (*Klra*, formerly *Ly49*) are expanded in the NKC. A common sign of these expanded gene families, along with allelic variation of members, is haplotypic variation in the number of genes and pseudogenes in populations/strains of same species (Marsch et al., 2003; Schenkel et al., 2013). The gene content of the LRC (human vs. primates) and NKC (mouse vs. rat) is known to vary even between closely related mammalian species and families; as a result, knowledge of NKR genes in a number of important species remains fragmentary or missing. Significant differences have been shown to exist within Artiodactyls, as for example between cattle and pigs (Sanderson et al., 2014; Schwartz et al., 2017; Schwartz and Hammond, 2018), but knowledge of the genes underlying camel NKR is lacking.

In the context of our work on the camelid immunogenome, the objective of this study was to characterize the genomic content of NKC and LRC with special focus on genes encoding natural killer cell receptors for MHC class I ligands in the two domestic camel species, *C. dromedarius* and *C. bactrianus*.

MATERIALS AND METHODS

Genomic Resources

Based on our new assembly CamDro2 of the *C. dromedarius* genome (Elbers et al., 2019), we characterized the NKC and LRC genomic regions and three natural cytotoxicity receptor genes (*NCR1*, *NCR2*, and *NCR3*). Their gene contents and organization were compared to the National Center for Biotechnology Information (NCBI) reference genomes for *C. dromedarius* (NCBI Genome accession code GCA_000767585.1) assembly PRJNA234474_Ca_dromedarius_V1.0, *C. bactrianus* (GCA_000767855.1) assembly Ca_bactrianus_MBC_1.0, and *Vicugna pacos* (GCA_000164845.3) assembly Vicugna_pacos-2.0.2. The selected orthologous genes were searched in two genomes of domesticated Artiodactyl species, cattle *Bos taurus* (GCA_000003055.5) assembly Bos_taurus_UMD_3.1.1 and pig *Sus scrofa* (GCA_000003025.6) assembly Sscrofa11.1. Various individual genomes for *C. dromedarius* (NCBI BioProject accessions: PRJNA269274, PRJNA269961, and PRJNA310822) and *C. bactrianus* (PRJNA183605 and PRJEB407) were searched in publicly available whole-genome shotgun contigs for candidate microsatellite markers. Likewise, individual whole genome sequencing reads for *V. pacos* (PRJNA233565 and PRJNA340289) were used to estimate single-nucleotide polymorphism (SNP) variability in selected genes of alpacas.

Annotation of Selected Genes in *C. dromedarius*

Alongside automatic computational annotation of genes in CamDro2 (see Elbers et al., 2019), selected unrecognized genes for NK receptors were manually annotated in the NKC and LRC genomic regions. First, we searched the *C. dromedarius* NCBI reference genome by tblastn algorithm of NCBI's BLAST®[1] for orthologous protein sequences to killer-cell lectin-like receptors recently identified in cattle as *KLR* genes (Schwartz et al., 2017). Second, the *ab initio* messenger RNA (mRNA) models complementary DNA (cDNA) for corresponding genes of all *KLR* gene lineages in the dromedary reference genome were inspected for completeness using NCBI's conserved domain database CDD search[2] and TMHMM Server v.2.0[3] for prediction of transmembrane helices in predicted proteins. The cDNA for *KLRE* was incorrect; thus, the cattle sequence (Schwartz et al., 2017) was used instead. These cDNA models were aligned against the CamDro2 chromosome 34 sequence using NCBI's Splign algorithm[4] and also BLAST®[1] searched in our full genomic assembly CamDro2. Identified genes were annotated accordingly. The Splign algorithm[4] was also used on scaffolds for the dromedary, Bactrian camel, and alpaca NCBI reference genomes.

The killer-cell immunoglobulin-like receptor *KIR* genes and leukocyte immunoglobulin-like receptor *LILR* genes were searched on CamDro2 chromosome 9 by the tblastn algorithm[1]. Individual immunoglobulin-like (Ig-like) domains and the cytoplasmic tail of Bactrian 3-domain receptor LILR (XP_010960360) served as query sequences. Orthologous and paralogous sequences were found. The corresponding genomic and predicted cDNA sequences were retrieved from the NCBI reference genomes for both camel species. The full length *KIR* and *LILR* cDNAs were cross-aligned with adequate scaffolds of reference genomes using Splign[4]. The predicted protein sequences were screened with CDD[2] and TMHMM Server v. 2.0 search[3]. The gene sequences were BLAST®[1] searched in CamDro2. Identified full-length genes were annotated, and gene fragments were recorded.

The natural cytotoxicity receptor genes *NCR1*, *NCR2*, and *NCR3* were BLAST® searched in the full assembly CamDro2 based on annotation of the dromedary NCBI reference genome. Although there exist numbers of alternatively spliced variants for each gene/protein, we focused on cDNA models of the longest variant (**Table 1**). The predicted dromedary, Bactrian camel, and cattle cDNA for *NCR3* were incomplete; therefore, we used the sequence predicted from the white-tailed deer *Odocoileus virginianus texanus* instead.

During the process of manual annotation, we also characterized the NKC and LRC regions, and comparisons were made between *C. dromedarius*, *C. bactrianus*, and *V. pacos*.

To allow comparisons with other studies and identification of orthologues, a standardized systematic nomenclature of NKR genes (Schwartz et al., 2017) was used. However, when referring to original reports on human or mouse genes, both the original and standard gene names and symbols were used.

Amplification and Next-Generation Sequencing

The gene-specific primers encompassing full-length genes with flanking sequences were designed on NCBI *C. bactrianus* gene sequences using Primer-BLAST[5] and checked for specificity against reference genomes of both camel species. The list of primer pairs used is available in **Supplementary Table 1**. Various compositions of PCRs and adequate thermal profiles used are summarized in **Supplementary Table 2**. PCR products were checked by 0.5% agarose gel electrophoresis and quantified by Invitrogen™ Qubit™ Fluorometer using Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). They were kept frozen at −20°C until massive parallel sequencing. Each individual's long-range amplicon of genes under study was indexed separately during library preparation using the NexteraXT DNA Library Preparation Kit (Illumina, San Diego, CA, USA) and sequenced on a MiSeq™ System (Illumina, San Diego, CA, USA) platform using the MiSeq™Reagent Kit v2 (500 cycles) according to manufacturer's protocol in different

[1] https://blast.ncbi.nlm.nih.gov
[2] https://www.ncbi.nlm.nih.gov/Structure/cdd
[3] http://www.cbs.dtu.dk/services/TMHMM
[4] https://www.ncbi.nlm.nih.gov/sutils/splign
[5] https://www.ncbi.nlm.nih.gov/tools/primer-blast

TABLE 1 | Number of predicted protein sequences differing by at least one amino acid residue.

Locus	mRNA model	Alias for protein	*Camelus dromedarius*	*Camelus bactrianus*	*Vicugna pacos*
KLRA	XM_010986205	Ly49	2	2	5
KLRC1	XM_010986199	NKG2A	3	3	5
KLRC2	XM_010986200	NKG2C	2	1§	5
KLRD	XM_010986196	CD94	3	1	3
KLRE	KX611578.1		2	1	5§
KLRI	XM_010986247		1	1	3
KLRJ	XM_010986202		2*	1*	5*
KLRK	XM_010986198	NKG2D	2	2	4
KIRDP	XM_011000160		NA	NA	NA
KIR3DL	XM_014563526	KIR3DL1, CD158	7*	3*	7
LILRB1	XM_010999297		3	1	7
LILRB2	XM_010961869+		6	2*	ND
LILRB3	XM_010961881+		4	5	ND
LILRA 2-Ig	XM_010961868		3	1	ND
LILRA 4-Ig	XM_011000157		3	4	ND
NCR1	XM_010961883	NKp46, CD335	2	3§	3
NCR2	XM_010973934	NKp44, CD336	2§	3	3
NCR3	XM_020897188	NKp30, CD337	2*	2*	3*

**Alleles predicted for a pseudogene; §Includes allele with premature stop codon; +Modified to represent all exons; NA, not applicable; ND, not determined.*

runs. The quality of the raw sequencing reads was checked using FastQC[6]. Low quality read ends were removed by Trimmomatic (Bolger et al., 2014) (SLIDINGWINDOW:4:15). Only reads longer than 150 bp were used for the mapping by BWA-MEM (Li, 2013). The alignment was post-processed using Samtools (Li et al., 2009) (sorting and conversions), GATK (DePristo et al., 2011) (indel realignment), and Picard[7] (PCR duplicates removal). Further, only mappings with the minimal mapped length of 70 bp, a maximum of 5% soft-clipping, and a maximum of 10% mismatches were kept using NGSUtils (Breese and Liu, 2013) and BBMap[8].

Microsatellite Markers

The repetitive sequences of di-, tri-, and tetra-nucleotides were searched in the NKC and LRC of the Bactrian camel NCBI reference genome by RepeatMasker[9]. Candidate microsatellites (msats) were identified by BLAST® search of repetitions flanked with 100 bp sequences in whole-genome shotgun contigs from three dromedaries and two Bactrian camels. The most diverse sequences with unique occurrence in genome were chosen, and primers were designed in OLIGO Primer Analysis Software v.4.0 (Molecular Biology Insights, Colorado Springs, CO, USA). Primer specificity was verified against the NCBI reference genomes of both camel species using NCBI's Primer-BLAST[5]. The PCR conditions were optimized for six msats, finalizing in one 5-plex (*CZM025-CZM029*) and one single (*CZM030*) PCR protocol. Reaction compositions were as follows: 1.0 µl 10 × Taq Buffer (Top-Bio, Prague, Czech Republic), 0.5 U CombiTaq DNA polymerase (Top-Bio, Prague, Czech Republic), 200 µM each dNTP (Thermo Fisher Scientific, Waltham, MA, USA), 0.1 µl of each primer of concentration 10 µM (**Table 2**), and 50 ng of genomic DNA. PCR reaction mix was supplemented with PCR grade H_2O (Top-Bio, Prague, Czech Republic) to a final volume of 10.0 µl. The thermal cycler ABI Verity 96 Well (Applied Biosystems, Foster City, CA, USA) was used for amplification. The thermocycling conditions consisted of initial denaturation at 95°C for 3 min; 30 cycles of denaturation at 95°C for 30 s, annealing at 56°C (64°C for *CZM030*) for 30 s, and elongation at 72°C for 30 s; and final elongation at 72°C for 60 min and holding at 7°C. All markers were then tested by fluorescent fragment analysis using Applied Biosystems® ABI PRISM 3500 and sized with GeneScan™ 500 LIZ® Size Standard (Thermo Fisher Scientific, Waltham, MA, USA). The data obtained from the fragment analyzer were evaluated using the GeneMapper® software v.4.1 (Thermo Fisher Scientific, Waltham, MA, USA).

Estimation of Genetic Variability in Camels and Alpacas

The dromedary DNA samples in this study were transferred from previous projects [Austrian Science Fund (FWF)P1084-B17 and P24706-B25; PI: P. Burger] and originated either from plucked hair, ethylenediaminetetraacetic acid (EDTA) blood collected commensally on Whatman FTA® cards (Sigma-Aldrich, Vienna, Austria) during routine veterinary controls, or from DNA extracts sent by collaborators under bilateral agreements. Samples were imported with permits from the Austrian Ministry of Labour, Social Affairs, Health and Consumer Protection. All the Bactrian camel samples were collected commensally during veterinary procedures for previous research projects (GACR 523/09/1972; PI: P. Horin). Details about the samples are provided in the **Supplementary Table 3**.

Selected genes of the NKC and LRC regions were genotyped by targeted resequencing of long-range PCR amplicons in both

[6]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[7]http://broadinstitute.github.io/picard/
[8]https://sourceforge.net/projects/bbmap/
[9]http://repeatmasker.org/cgi-bin/WEBRepeatMasker

TABLE 2 | Characteristics of microsatellite markers in natural killer complex (NKC) and leukocyte receptor complex (LRC) regions.

Marker	Type of repetition	*C. dromedarius* amplicon size	Forward primer Reverse primer
CZM025	(AC)n	184–210 bp	6-FAM-CCCACAGGCTGTTCCTCAAA-3′ 5′-TGTCCTGGTTATGGGAAGATGG-3′
CZM026	(GT)n	216–236 bp	NED- CTTCCAAATGCACCTGAAACATC-3′ 5′-TGACTTCAAGGGAATGCCTCAA-3′
CZM027	(TG)n	233–247 bp	6-FAM-GGCTGGACAGTGCAAATTTTACC-3′ 5′-GCACCATCTCTGGAGGCTAAGAG-3′
CZM028	(AC)n	95–113 bp	NED- TAATACTCGCCATCCTTCTGCCT-3′ 5′-GAGACCCTCCGGTGTAGAAAGC-3′
CZM029	(AC)n	169–193 bp	NED- ATCATGTCAGCATTGCTTTGGAA-3′ 5′-CATGTGTCCTGACGCTGGAA-3′
CZM030	(CA)n	167–187 bp	6-FAM-CCGTCAGCTGGAAATTTGTCTCT-3′ 5′-AGAGTCAGGAGGGCTTCTAGGCTA-3′

camel species. For comparison, individual genotypes in the same batch of genes except *LILR* genes were acquired for alpacas by data mining.

Two panels of 10 animals were created from collections of samples originating from various populations. The *C. dromedarius* panel encompassed individuals coming from Jordan (Irbid), Iran, Saudi Arabia (Magaheem and Wadda), Canary Islands, UAE (Dubai), Kenya, Sudan, Nigeria, and Kazakhstan. The genomic DNA was previously isolated by phenol-chloroform extraction and kept frozen at −20°C. The C. *bactrianus* panel consisted of individuals from three Mongolian regions (Bayan Ovoo, Galshar, and Norovlin). The genomic DNA was isolated from frozen archived whole blood samples by NucleoSpinBlood© Kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's protocol. The genes of interest were isolated by PCRs on genomic DNA, obtained amplicons were indexed to track individual samples, and then were sequenced in multiple Illumina next-generation sequencing (NGS) runs and mapped to adequate reference sequence for amplicon (see above).

A panel of four alpacas was created from publicly available whole-genome sequencing projects. The raw data of Illumina NGS runs were downloaded from the European Nucleotide Archive database[10] (ENA accession numbers SRR1552593-1552609, SRR4095109, SRR4095110, and SRR4095135). The quality was checked using FastQC[6] and Kraken package (Davis et al., 2013). Adapter and quality (Phred < 15) trimming was performed by Cutadapt (Martin, 2011). BWA-MEM (Li, 2013) was used for the alignment, and the alignments were post-processed by Samtools (Li et al., 2009) (sorting and conversions), GATK (DePristo et al., 2011) (indel realignment), and Picard[7] (PCR duplicates removal). Alignments were further filtered using NGSUtils (Breese and Liu, 2013) and BBMap[8] for maximum soft-clipping (5%), maximum number of mismatches (10), minimum mapped length (35 bp), maximum soft-clipping (5%), and minimum mapping quality (MAPQ 40). The reference sequences for mapping were retrieved from the *V. pacos* NCBI reference genome. Most *V. pacos* sequences were framed by the primer sequences used in camels.

[10]https://www.ebi.ac.uk/ena

Data Analysis

The alignments of reads to the reference sequence were inspected using IGV software[11]. The variable positions (variant in homozygous state) and confirmed sequence variants (variant detected in heterozygous state) were treated as SNPs. They were written to consensus sequences for each animal using IUPAC nucleotide ambiguity codes in BioEdit, version 7.2.6. (Hall, 1999) along with insertions/deletions, and sequences from same animal species were manually aligned. The number of SNPs was counted using DnaSP version 5.10 program (Librado and Rozas, 2009), and frequency was calculated as percentage. The cDNA sequences were *in silico* extracted in BioEdit v.7.2.6, based on mRNA models for each gene (**Table 1**) and checked by Splign[4] for completeness. Haplotypes of each diploid individual were reconstructed for every panel and gene (cDNA) separately using PHASE (Stephens and Donnelly, 2003) algorithm in DnaSP v.5.10. The coding sequences were extracted in BioEdit v.7.2.6. The number of SNPs was counted in DnaSP v.5.10, and the percentage of coding sequence length calculated.

Amino acid sequences were deduced from coding sequences in BioEdit v.7.2.6. The manually aligned predicted protein sequences were compared. Sequences differing in at least one amino acid position were numbered and designated as different alleles of a particular gene.

A phylogenetic analysis of sequences obtained by long-range PCR or data mining was done separately for NKC (C-type lectin-like) and LRC (immunoglobulin-like) genes. The nucleotide coding region sequences were aligned by ClustalW Multiple alignment algorithm in BioEdit v.7.2.6. One haplotype per gene was chosen for each species to represent the respective loci of the dromedary, Bactrian camel, and alpaca. Corresponding cattle and pig sequences retrieved from NCBI's GenBank[12] were used for a comparison. The maximum likelihood phylogenetic trees were constructed in MEGA5 (Tamura et al., 2011) based on the Tamura 3-parameter model and the partial deletion method (95% cutoff) with 100 bootstrap repetitions (Tamura, 1992).

[11]https://software.broadinstitute.org/software/igv/

[12] https://www.ncbi.nlm.nih.gov/genbank/

RESULTS

The general organization of the two genomic regions, the natural killer complex (NKC) and the leukocyte receptor complex (LRC), containing genes and gene families encoding the NK cell receptors annotated based on the dromedary genome assembly CamDro2, was established and is represented in **Figure 1**. The phylogenetic trees of the genes analyzed are shown in **Figures 2** and **3** for NKC and LRC genes, respectively. A summary of the allelic variants of the predicted proteins for the genes genotyped in dromedaries and Bactrian camels is given in **Table 1**. The alignments of amino acid sequences with depictions of their protein domains are provided in the **Supplementary Figures S1–S3**. An overview of SNPs of the full genes and coding sequences is found in **Supplementary Table 4**. Protein homology of selected receptors in dromedary relative to Bactrian camel, alpaca, cattle, and pig orthologues is summarized in **Supplementary Table 5**.

Natural Killer Complex

The NKC region encompassing approximately 0.9 Mbp was localized on chromosome 34 of CamDro2. Twenty-six genes encoding receptors with the C-type lectin-like domain (CTLD) of different lineages were identified in this region. No expansion of any *KLR* gene family was observed in the dromedary genome. Most *KLR* genes clustered at one end of the region. This cluster contains five functional genes (*KLRA*, *KLRD*, *KLRE*, *KLRI*, and *KLRK*), two functional members of the *KLRC* family, and two pseudogenes (*KLRH* and *KLRJ*). *KLRC* is the only gene family with two members sharing the same CTLD but signaling oppositely: *KLRC1* codes for an inhibitory receptor, while *KLRC2* encodes an activating receptor. Members of three families (*KLRB*, *KLRF*, and *KLRG*) are located at the opposite end of the NKC region, separated from each other by a group of C-type lectin (*CLEC*) genes. Two members of each of these families were found in the CamDro2 dromedary genome. The genes *KLRB1* and *KLRB1B* have the standard structure of inhibitory receptors with a cytoplasmic tail containing an immunoreceptor tyrosine-based inhibitory motif (ITIM). Genes encoding their putative ligands (*CLEC2D* and *CLEC2F*, respectively) were found in the vicinity. Similarly, the genes *KLRF1* (NKp80) and *KLRF2* (NKp65) encoding activating receptors are located in close proximity of genes coding for their predicted ligands, AICL and KACL (*CLEC2B* and *CLEC2A*, respectively). *KLRG1* encoding an inhibitory receptor marks the boundary of the NKC. *KLRG2* was found outside of NKC, on chromosome 7.

While in the *C. dromedarius* NCBI reference genome the NKC is split into two scaffolds (NW_011591409, NW_011591669), it is contained within a single scaffold (NW_011511552) in the *C. bactrianus* NCBI reference genome. The gene content and gene orientations are the same in both genomes. The only exception is the presence of a premature stop codon in the Bactrian *KLRC2* sequence, which thus seems to be a pseudogene.

Variability of NKC Receptors

Since no expansion of *KLR* gene families was observed, we focused on the allelic variation of inhibitory receptors supposed to recognize MHC class I ligands. Due to their poor quality, some samples were not successfully amplified by long-range PCRs. Because of an apparent mixed ancestry (*C. dromedarius* X *C. bactrianus*) of one *C. dromedarius*, heterozygous sequences of mixed origin were removed. Consequently, different numbers of genotypes were retrieved for different genes as indicated in **Supplementary Table 4**. Despite polymorphisms existing in the genomic and predicted mRNA sequences (**Supplementary Table 4**), none of the tested genes were found to have more than three protein variants. Five of the eight tested genes in *C. bactrianus* were monomorphic on the protein level.

KLRA encodes an inhibitory receptor with one ITIM signaling motif in its cytoplasmic tail and a relatively long extracellular stalk region (over 70 amino acids). Two variants of this receptor molecule were predicted in *C. dromedarius*, sharing the same CTLD but differing by one amino acid residue in the cytoplasmic tail. A KLRA variant with the same CTLD occurs frequently in *C. bactrianus*. A second variant of Bactrian KLRA differs by nine amino acids (one in the cytoplasmic tail, four in the stalk, and four in the CTLD).

KLRC1 codes for an inhibitory receptor with two ITIMs. Three variants of the KLRC1 protein were identified in each camel species. One KLRC1 variant was shared by both camel species, and two additional variants in each species differed by only one amino acid. Only two different CTLD variants were present in each species.

KLRC2 codes for an activating receptor with a charged amino acid residue (lysine) in the transmembrane domain. In *C. dromedarius*, *KLRC2* appears to encode two variants of a functional receptor. In *C. bactrianus*, this gene seems to be monomorphic with a premature stop codon shortly after the origin of translation.

The *KLRD* gene product consists of a CTLD with a short stalk and a transmembrane domain with no signaling motif. For KLRD, one protein sequence common to both camel species was observed, with two additional variants found in *C. dromedarius*, differing by one amino acid each. All three camel genes, *KLRC1*, *KLRC2*, and *KLRD*, have codons for cysteine residues in the stalk region of the protein, allowing formation of disulfide links and of heterodimers KLRD/KLRC1 and KLRD/KLRC2.

Another pair of presumably interacting receptors forming noncovalent heterodimers is KLRE/KLRI. The striking features of Old World camelid *KLRE* are the presence of sequence for an ITIM in the cytoplasmic tail of the protein and the existence of a second variant in *C. dromedarius* with a duplication of six amino acid residues in the CTLD. In *C. bactrianus*, *KLRE* encodes only one variant of the protein sequence.

The *KLRI* gene showed very limited polymorphism in both camel species at the genomic level, encoding only one protein variant with only one functional ITIM (and one mutated motif) in the cytoplasmic tail of the molecule.

The predicted protein product of the *KLRJ* gene was identical in both camel species. Another sequence variant in the dromedary camel differed by only one amino acid in the stalk region. According to the adopted mRNA model, all these sequences contain a premature stop codon in the CTLD of the protein.

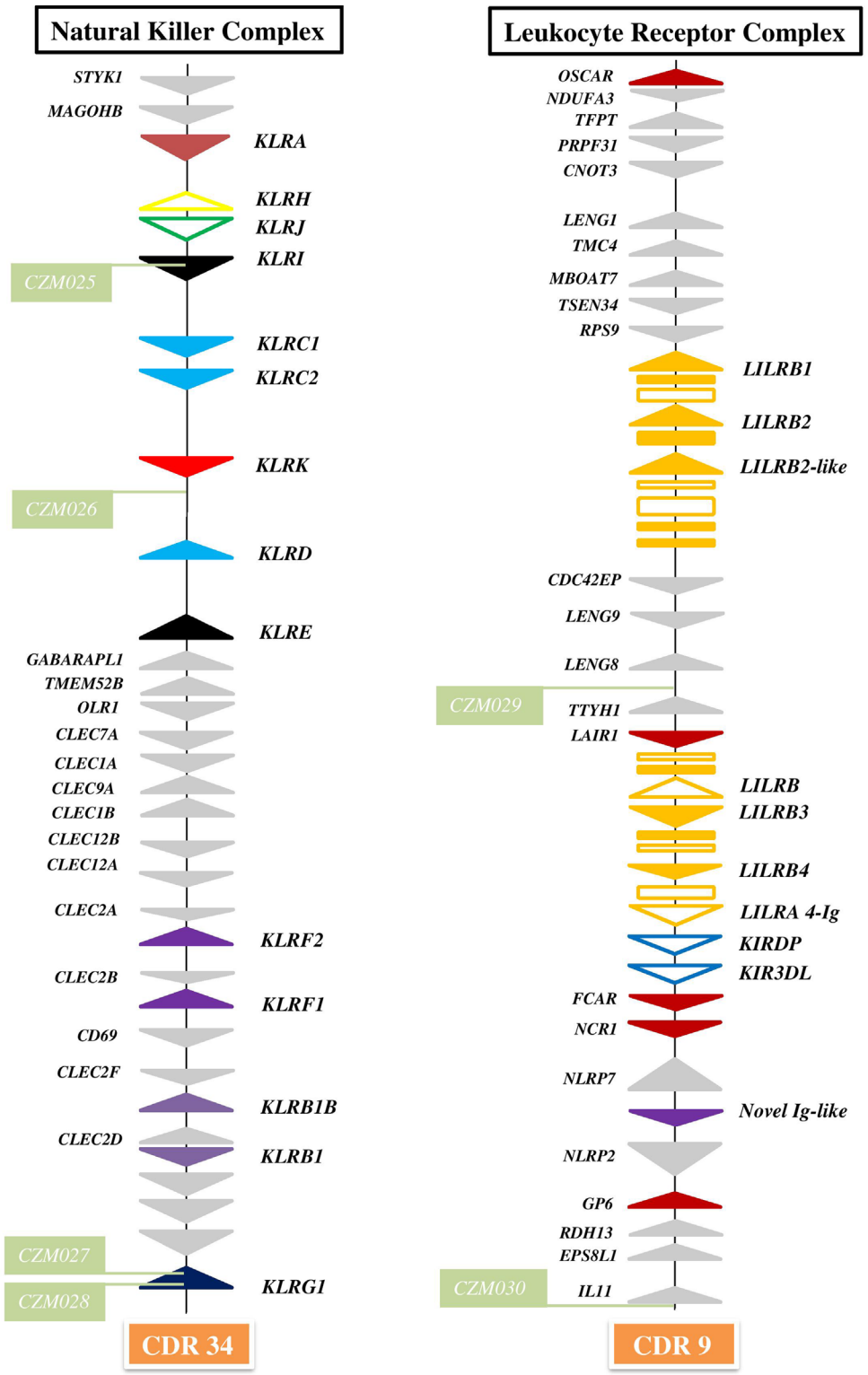


FIGURE 1 | Organization of genomic regions encoding NK cell receptors in dromedary camel. The NKC was delineated by *STYK1* and *KLRG1* genes on chromosome 34 (CDR34) of CamDro2 between 11.61 and 12.50 Mb. *KLR* genes are represented as *solid color triangles*, *KLR* pseudogenes as *empty color triangles*, and lectin-like *CLEC* or non-lectin genes as *solid grey triangles*. The LRC was found between the *OSCAR* and *IL11* genes on chromosome 9 (CDR9) in the region 63.00–72.01 Mb. *LILR* genes are represented as *solid orange triangles*, *LILR* pseudogenes as *empty orange triangles*, immunoglobulin-like domains or cytoplasmic domain gene fragments as *orange rectangles*, *KIR* pseudogenes as *empty blue triangles*, other types of Ig-like genes as *solid color triangles*, and different types of flanking genes as *solid grey triangles*. *Green rectangles* mark positions of newly developed microsatellite markers *CZM025–CZM30*.

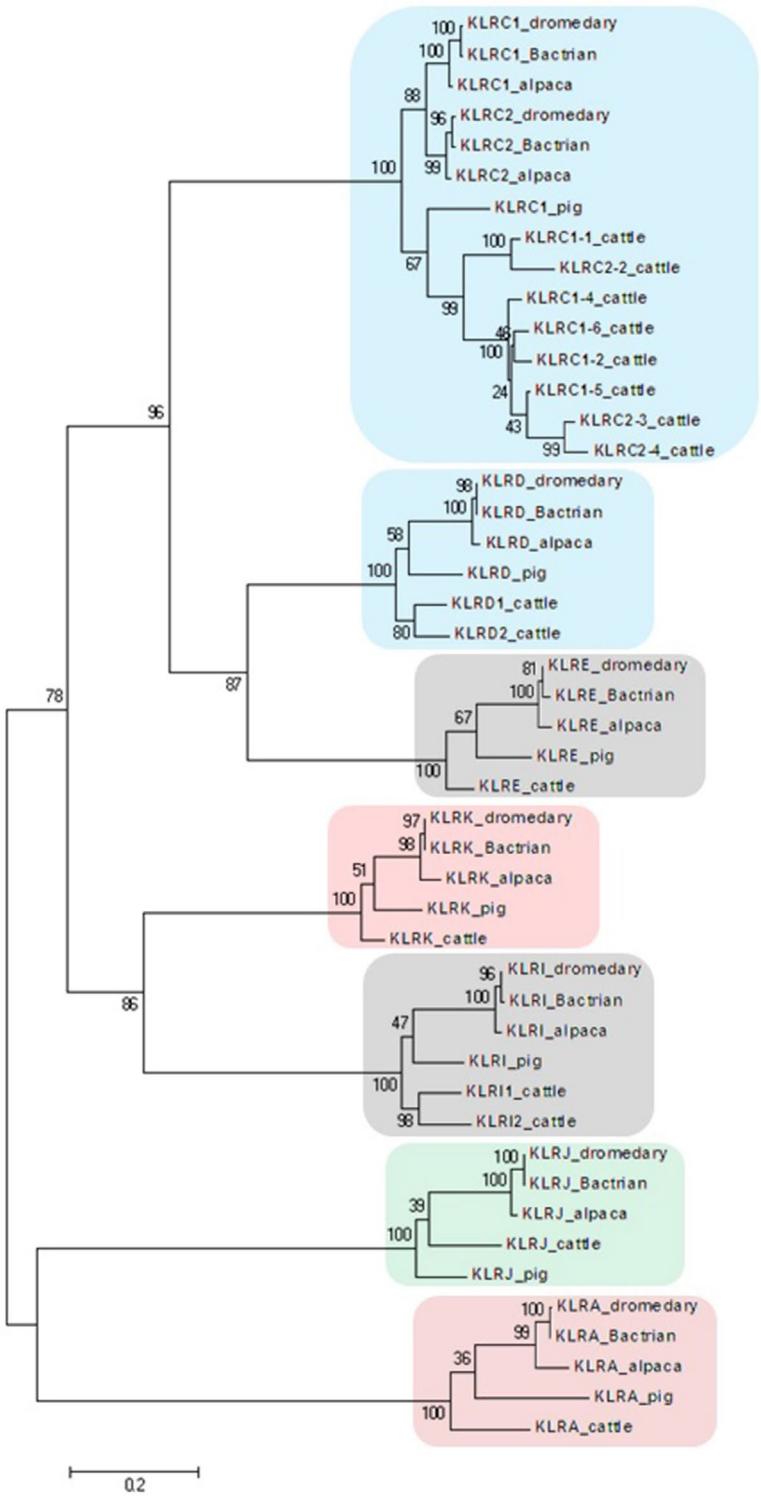


FIGURE 2 | Phylogeny of nucleotide coding sequences for NKC C-type lectin-like genes analyzed by long-range PCR/data mining. The percentage of trees (out of 100 bootstrap replicates), in which the associated sequences clustered together is given at branch nodes. Branch lengths are expressed as the number of substitutions per site. Clusters of genes are highlighted according to the color scheme used in Figure 1. Haplotypes generated in this study were chosen (one per gene) to represent loci of *Camelus dromedarius* (dromedary), *Camelus bactrianus* (Bactrian), and *Vicugna pacos* (alpaca). Comparisons were made to *Bos taurus* (cattle) sequences retrieved from GenBank (accession numbers: KX611576, KX611577, KX611578, KX698607, NM_174376.2, NM_001075139.1, NM_001098163.1, and NM_001168587.1) and *Sus scrofa* (pig) sequences (accession numbers: NM_213813.2, NM_214338.1, XM_005655677.3, XM_005655679.3, XM_013988381.2, XM_013988416.2, and XM_021092357.1).

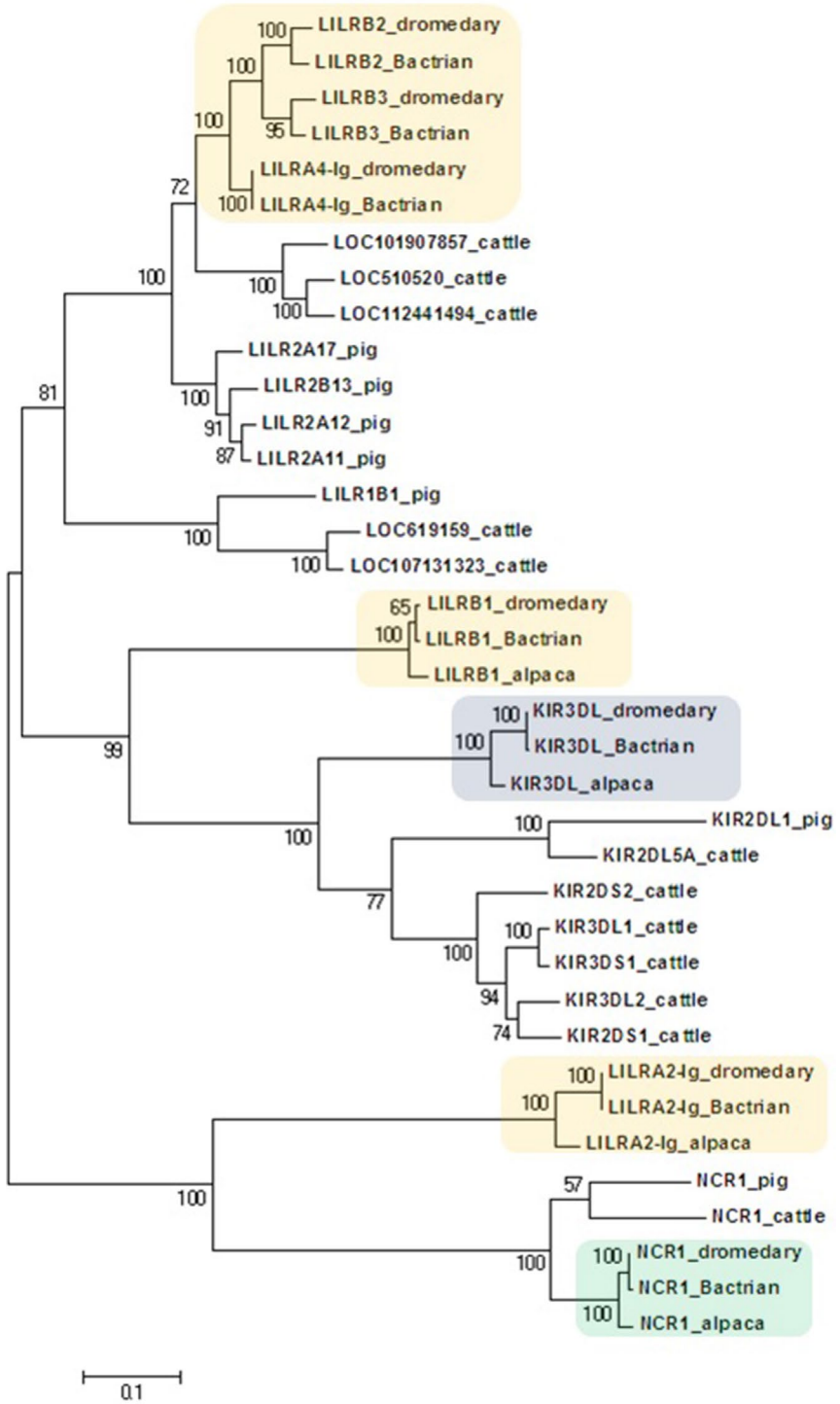


FIGURE 3 | Phylogeny of nucleotide coding sequences for LRC immunoglobulin-like genes analyzed by long-range PCR/data mining. The percentage of trees (out of 100 bootstrap replicates) in which the associated sequences clustered together is given at branch nodes. Branch lengths are expressed as the number of substitutions per site. Haplotypes generated in the study were chosen (one per gene) to represent loci (*colored*) of *Camelus dromedarius* (dromedary), *Camelus bactrianus* (Bactrian), and *Vicugna pacos* (alpaca). The nomenclature for the camelid *LILR* gene family is provisional and will change when a complete assembly of this region is available. Some alpaca *LILR* genes were not included due to an incomplete resolution of this family in the reference genome. The alpaca's *LILRA 2-Ig* sequence was retrieved from GenBank (accession number XM_015252448.1). A selection of *Bos taurus* (cattle) sequences (accession numbers: NM_174740.2, NM_181451.1, NM_183365.1, NM_001008415.1, NM_001097567.1, NM_001098089.1, XM_005201067.4, XM_024978801.1, XM_024978806.1, XM_024978818.1, XM_024978824.1, and XM_024978827.1) and all functional *Sus scrofa* (pig) sequences (accession numbers: NM_001113218.1, NM_001123143.1, NM_001128451.1, XM_003134173.4, XM_013998762.2, XM_021094960.1, and XM_021094977.1) were used for comparison.

KLRK in both camel species codes for a functional activating receptor with a charged amino acid (arginine) in the transmembrane region. Two proteins with variant CTLD were recognized in *C. dromedarius*, while two proteins in *C. bactrianus* share the same CTLD as one of the dromedary KLRK.

In the phylogenetic trees obtained, all NKC genes sequences of both camel species clustered with their putative orthologues in alpaca, cattle, and pig (**Figure 2**).

Leukocyte Receptor Complex

The LRC region of approximate length 0.7 Mbp was localized to chromosome 9 of CamDro2. Fifteen full-length genes encoding receptors containing immunoglobulin-like (Ig-like) domains of various lineages were identified in the LRC region. Besides two *KIR* pseudogenes containing ITIM domains and an expanded *LILR* gene family, a singular Ig-like receptor was found in the vicinity of *FCAR* and *NCR1*, located between *NLRP7* and *NLRP2*. It comprises two Ig-like domains different from those of the *LILR* and the *KIR* genes and has a long cytoplasmic tail with two ITIMs. It is a novel type of LRC gene, similar to a gene recently identified in pigs (Schwartz and Hammond, 2018). Based on its structure, this inhibitory type of receptor gene may be functional, similarly to pigs.

The expanded family of *LILR* genes is organized in two distinct clusters. The first region spanning approximately 141 kb is located between the genes *RPS9* and *CDC42EP5*. This region contains three putatively functional genes, *LILRB1*, *LILRB2*, and a *LILRB2-like* sequence. *LILRB1* codes for an inhibitory receptor with three Ig-like domains. *LILRB2* and *LILRB2-like* each encode four Ig-like domains and a cytoplasmic domain with ITIMs. Several fragmented sequences containing Ig-like domains were identified within this region as well. The second *LILR* region spanning approximately 127 kb is located between *LAIR1* and the two *KIR* pseudogenes. Three full-length *LILR* genes and a pseudogene were identified in this region. *LILRB3* codes for an inhibitory receptor; it comprises four Ig-like domains, a transmembrane region, and a cytoplasmic tail with two intact ITIMs. *LILRB4* also codes for an inhibitory receptor, but with only two Ig-like domains. In addition, there are a potentially functional activating *LILRA* gene and a *LILRB* pseudogene (containing an ITIM sequence) located in the same region. The predicted LILRA contains four Ig-like domains in its extracellular region but has no signal peptide sequence. The cytoplasmic domain is short and contains no ITIMs. Several fragments with complete or partial Ig-like domains were also found in this region.

In the current *C. dromedarius* NCBI reference genome, the LRC is split amongst at least four scaffolds (NW_011593473, NW_011591120, NW_011595380, and NW_011591711). They matched our CamDro2 assembly in terms of the number and orientation of orthologous non-Ig-like genes recognized by automatic annotation. Single Ig-like genes and two *KIR* pseudogenes, but not expanded *LILR* genes, were unraveled. The Bactrian LRC of NCBI reference genome is contained within a single scaffold (NW_011515311), but assembly of the expanded *LILR* genes is not resolved and *LILRB2-like* is missing. The overall LRC organization in the *C. bactrianus* reference genome is the same as that of the LRC of dromedaries (**Figure 1**).

Variability of LRC Receptors

Based on PCR and resequencing of representative panels of *C. dromedarius* and *C. bactrianus*, individual genotypes could be successfully identified for most of the genes analyzed. However, the amplification of *KIR3DL* sequences in Bactrian camels provided only limited numbers of sequences (**Supplementary Table 4**). Similar to NKC genes, some sequences from one *C. dromedarius* individual were removed due to their mixed origin.

The *KIR3DL* gene contains a 2-bp deletion in the exon for the third Ig-like domain, causing a frameshift and a premature stop codon. This deletion is identical in both camel species. The locus *KIRDP* contained sequences with premature stop codons and frameshift mutations in both camel species. The same was found in NCBI reference genomes.

Therefore, we assigned *KIRDP* and *KIR3DL* sequences provisionally as pseudogenes in both species. In contrast to the low polymorphism of the NKC receptors, higher numbers of variable amino acid sites were found within KIR3DL. An *in silico* 2-bp insertion resulted in three and seven predicted full-length protein variants in Bactrian camels and dromedaries, respectively (**Table 1**).

LILRB1 encodes a protein of similar structure to *KIR3DL*, with three constant-type Ig-like domains and two functional ITIMs in its cytoplasmic tail. Unlike other members of the *LILR* family coding for receptors with four Ig-like domains, *LILRB1* has no variable-type Ig-like domain between its first and second domains. Nevertheless, in *C. dromedarius*, only three variants with minor changes in their amino acid sequences were found, and in *C. bactrianus*, this gene appears to be monomorphic.

The gene *LILRB2* of *C. dromedarius* encodes a functional inhibitory receptor with four Ig-like domains and two ITIMs. Six variants recognized in the panel of dromedaries share the same Ig-like domains and differ only by two amino acids in the transmembrane region and one in the cytoplasmic tail. All *LILRB2-like* sequences obtained by PCR were identical to *LILRB2* sequences retrieved from the same dromedaries' DNAs. In *C. bactrianus*, *LILRB2* encodes two mRNAs with a premature stop codon in the first Ig-like domain and differs by 37–38 amino acid positions from its dromedary counterparts. No PCR products were obtained for *LILRB2-like* from the Bactrian camel panel.

LILRB3 encodes a receptor with 82% identity (86% similarity) to LILRB2 in *C. dromedarius*. All four protein variants had only two different Ig-like domains (the second and third) with one amino acid change each. Five variants of the Bactrian LILRB3 had 16 inter-species specific positions with another 11 amino acids differing within species.

Sequences of two activating *LILRA* genes were retrieved by PCR. One of them, containing two Ig-like domains, arginine in the transmembrane region, and a long cytoplasmic tail, but with the first ITIM deleted and the second mutated, was provisionally named as *LILRA 2-Ig*. Both camel species shared one variant of the LILRA 2-Ig protein, and two additional variants were present in the dromedary. The second activating gene was designated

LILRA 4-Ig as it contained four Ig-like domains, arginine in its transmembrane region, and a short cytoplasmic tail with no signaling motif. Three variants of the dromedary and four variants of the Bactrian camel LILRA 4-Ig protein are very similar, differing in only six amino acid positions. One specific variant with an in-frame deletion of 25 amino acids in the third Ig-like domain was shared between species.

The phylogenetic tree constructed for the coding regions of the camel LRC genes (**Figure 3**) and their homologs in alpaca, cattle, and pig revealed three main clusters of genes characterized by the overall structure of the encoded receptor. The first cluster grouped genes with four Ig-like domains receptors (LILRs). The second group was a cluster of genes coding for receptors with three or two Ig-like domains (KIRs and LILRB1). The third cluster was formed by genes encoding receptors with two Ig-like domains (NCR1 and LILRA2-Ig). Within the first cluster, three distinct camel genes, *LILRB2*, *LILRB3*, and *LILRA4-Ig*, clustered with various cattle and pig *LILR* genes. Likewise, various cattle and pig *KIR* genes formed a cluster with camelid *KIR3DL* genes. This cluster was related with the cluster of camelid *LILRB1* genes, while the cluster containing *NCR1* gene sequences was related to the camelid *LILRA2-Ig* genes. As no cattle and/or pig homologs of camelid *LILRB1* and *LILRA2-Ig* genes could be identified in the reference genomes of these two species, they did not appear in the trees constructed.

The Natural Cytotoxicity Receptors

The predicted proteins of the *NCR1*, *NCR2*, and *NCR3* genes were studied as potentially activating immunoglobulin-like receptors for various ligands different from MHC class I.

The *NCR1* gene is located within the LRC, and the structure of camelid NCR1 is similar to the structure of this gene's products in other species, with two extracellular Ig-like domains and a charged residue (arginine) in the transmembrane domain, which allows its interaction with activating adaptor proteins. Two allelic variants were identified in each camel species that differed from each other by only one or two amino acids. *C. bactrianus* possessed one additional variant containing a premature stop codon in the first Ig-like domain.

The *NCR2* gene is located on chromosome 20 of CamDro2. It encodes a functional receptor with one extracellular Ig-like domain and a charged residue (lysine) in the transmembrane domain. One allelic variant of the receptor is shared by both camel species. Another variant, found only in *C. dromedarius*, has a premature stop codon in the stem region of the putative molecule. Two additional variants were identified in *C. bactrianus*, differing by one amino acid each.

The *NCR3* gene is also located on chromosome 20, within the MHC region. All sequences in both camel species contained the same two premature stop codons; this gene thus seems to be nonfunctional in camels.

"*In Silico*" Comparison With *V. pacos*

The alpaca is evolutionarily the most closely related species to the Old World camels. The *V. pacos* NCBI reference genome contains two scaffolds (NW_005882720 and NW_005883060) with *CLEC* and *KLR* genes. The gene content and organization of the alpaca NKC region was found to be very similar to that of the camel NKC, with similarities of amino acid sequences ranging from 88% to 98%. Based on publicly available alpaca genomes, the extent of polymorphism of genomic as well as of protein sequences was higher than in either of the camel species. Five protein variants were predicted for *KLRA*, *KLRC1*, *KLRC2*, and *KLRE*. The amino acid changes were concentrated in the CTLD and the stem of KLRA and in the CTLD in KLRC2. In contrast, they were evenly distributed throughout KLRC1. *KLRE* coded for four functional protein variants with only three different CTLDs, and amino acid changes were concentrated mostly in the cytoplasmic tail. One allele had a 1-bp insertion leading to frameshift and a premature stop codon in the CTLD part of the receptor. The same ITIM motif as in camels was recognized in all five variants. The three variants of KLRD identified differed by one amino acid each in the stem region of the receptor and thus shared the same CTLD. *KLRI* also coded for three variants with the same CTLD, differing only in the cytoplasmic tail. Due to a non-synonymous substitution, a second ITIM was recreated in one of the variants. The five protein variants predicted for *KLRJ* differed in their CTLDs, although all of them contained the same premature stop codon according to the mRNA model. *KLRK* haplotypes coded for a potentially functional activating receptor (arginine in the transmembrane region). Only one amino acid difference was observed in two CTLD types shared by the four KLRK protein variants.

The alpaca LRC region is fragmented amongst numerous scaffolds of the NCBI reference genome. One of them (NW_005882947) contains a four-domain *LILRB* gene, *KIRDP*, *KIR3DL*, *FCAR*, *NCR1*, *GP6*, and a novel Ig-like gene, while another scaffold (NW_005883177) contains *LILRB1*, comprising three Ig domains, and *LILRB*, with four Ig domains. Only fragments of *LILR* genes could be found in the rest of the relevant scaffolds. Like in camels, *KIRDP* contains various frameshifts. In contrast, *KIR3DL* has retained an intact genomic sequence and is thus likely to encode a functional inhibitory receptor, but with only one functional ITIM. The second ITIM is mutated in all protein variants. Seven variants of KIR3DL were identified, differing in 11 amino acid positions in total. These sites are equally distributed throughout the molecule. *LILRB1* also codes for a potentially functional inhibitory receptor with three Ig-like domains and seven identified variants. They differed in 17 positions located in two of the Ig-like domains, the stem and the cytoplasmic tail. The protein homology in comparison with dromedary camel *KIR3DL* and *LILRB1*counterparts reached 93%.

The gene *NCR1* of the alpaca encodes a functional activating receptor with a charged amino acid residue (arginine) in the transmembrane region. The three detected allelic variants differed only in the stem and transmembrane regions. *NCR2* codes for a functional activating receptor (lysine in the transmembrane region) as well. Three protein variants with minor changes in the signal peptide and the cytoplasmic tail not affecting their potential function were identified. All identified *NCR3* sequences have premature stop codons in the Ig-like domain. The amino acid sequence similarity to camel sequences was 96%.

DISCUSSION

In contrast to the rather conservative organization of the mammalian major histocompatibility complexes, natural killer cell receptor genes and their complex genomic regions are evolutionarily flexible. Several different types of genomic organization of the NKR regions have been recognized in mammals (Martin et al., 2002; Hao et al., 2006; Guethlein et al., 2015), and sometimes striking differences have been observed between related taxa (Kelley et al., 2005; Sanderson et al., 2014; Schwartz et al., 2017; Schwartz and Hammond, 2018). Therefore, studies of genes encoding NK cell receptors may contribute to our understanding of the heterogeneity of NK cell functions in particular mammalian species. At the same time, these genes and especially complex genomic regions such as NKC and LRC represent a relevant model for evolutionary biology. Characterization of NKR genes in so far poorly studied species and/or families can bring new information on evolutionary mechanisms governing this part of the mammalian immunogenome. Despite the importance of camels as a model for immunogenetic studies (Ciccarese et al., 2014), virtually nothing is known about NK cells in camels in terms of their morphology, functions, their surface receptors, and/or underlying genes. In this context, this study represents the first report on the NKC and LRC genomic regions and on *NCR* genes in Old World camels and their comparison with a New World camelid, the alpaca.

Whole genome sequences of Old World camels, *C. dromedarius*, *C. bactrianus*, and *Camelus ferus* (Jirimutu et al., 2012; Wu et al., 2014; Fitak et al., 2016) and of the alpaca *V. pacos* (NCBI Genome[13] accession GCA_000164845.3) are currently available. However, their annotation is rather fragmentary and largely composed of predicted sequences generated *in silico*, based on homologies and sequence similarities with other mammalian species. Taking into consideration the quality of resources including the availability of biological material, we focused on the two domestic camel species, *C. dromedarius* and *C. bactrianus*. Even for these species, however, the current status of whole genome assemblies proved to be insufficient for a correct annotation of NKR genes, especially of the LRC, containing multiple copies of sometimes highly similar sequences and exhibiting copy number variations. Therefore, the major resource for our analyses was a new genome assembly of the *C. dromedarius* genome obtained by a combination of several long-read sequencing techniques and bioinformatic approaches (Elbers et al., 2019).

In all genes selected for sequence analyses, the genomic sequences of NKR genes were highly similar in both camel species studied as well as to available alpaca sequences. Such a high sequence similarity was observed for a number of other genes and was also characteristic for MHC class II sequences (Plasil et al., 2016). Therefore, it seems that PCR failures observed in some cases probably do not indicate polymorphisms in the primer binding site. Taking also into consideration the generally low polymorphism of the camel genomes (Fitak et al., 2016), the occurrence of a putative polymorphic variant on both chromosomes is not too likely. Moreover, the PCR failures concerned mostly the loci *KLRC2* and *KIR3DL* in the Bactrians, which seem to be both pseudogenes, so we have not explored them further for the purposes of this study. Nevertheless, both loci merit to be further investigated in the future. Information that the monomorphic status of the Bactrian *KLRC2* could be explained by allelic drop-out or existence of copy number variation, i.e., partial/total deletion of *KLRC2* from some Bactrian NKC haplotypes, and that polymorphic amino acid positions within the *KIR3DL* sequences were concentrated in the second immunoglobulin-like domain, known to interact with MHC class I ligands in mammals and in the stalk region of the molecule, need to be explored.

Another potential technical problem is the use of long-range PCRs for amplifying related members of a gene family, which may produce chimeric products. The NKC genes analyzed here were in majority single (not duplicated) genes with characteristic sequences. The LRC genes analyzed, with only *LILR* genes as members of expanded gene family/families, were different to such an extent that we could distinguish them. In addition, both types of phylogenetic trees clearly supported the individuality of each gene. Moreover, sequences successfully amplified as a whole matched the reference sequences. The remaining ones were amplified in two pieces, and again, they matched the reference sequences. Although we have checked the overlapping sequences of two-piece PCRs and they did not indicate more polymorphisms, we cannot strictly exclude the possibility that such a sequence could be composed of pieces of two different yet highly homologous loci, taking also into consideration numerous fragments of *LILR* genes observed in CamDro2.

The genomic organization of the NKC region of camelids differs from cattle, the phylogenetically most closely related species, whose NKR genes have been studied so far. While in cattle an expansion of *KLRC* and *KLRH* genes was reported (Schwartz et al., 2017), the minimal set of *KLR* genes observed in camelids resembles the genomic organization of the NKC of the domestic pig. Similarly, the leukocyte receptor complex of camels is strikingly different from the cattle LRC containing expanded *KIR* (Sanderson et al., 2014) and *LILR* genes (Hogan et al., 2012). In camels, the LRC with non-expanded *KIR* genes and several pseudogenes seems to be even less complex than in the pig (Schwartz and Hammond, 2018).

Within the natural killer complex, all types of *KLR* genes identified in mammals (Hao et al., 2006) were found. None of them apparently expanded into a large family; the maximum number of members within a family was two. Similar to other mammalian species, the *KLRA* gene codes for a homodimeric type II inhibitory receptor (Ly49) (Dimasi and Biassoni, 2005), *KLRC1* encodes an inhibitory receptor with two ITIMs motifs (NKG2A) (Vance et al., 1998), *KLRC2* encodes an activating receptor (NKG2C) (Lanier et al., 1998), and *KLRG1* codes an inhibitory receptor for cadherin molecules (Ito et al., 2006). These data suggest that the function of these molecules could be very similar to human and other mammalian NK receptors, especially in terms of their capacity to form heterodimers CD94/NKG2A (KLRD/KLRC1), CD94/NKG2C (KLRD/KLRC2)

[13]https://www.ncbi.nlm.nih.gov/genome

(Braud et al., 1998), and/or KLRE/KLRI (Saether et al., 2008). In humans, the heterodimers CD94/NKG2A (KLRD/KLRC1) and CD94/NKG2C (KLRD/KLRC2) recognize a relatively low polymorphic non-classical MHC class I ligand HLA-E, and their polymorphism is also rather low (Braud et al., 1998). Contrary to rats, in which KLRH recognize MHC class I ligand (Daws et al., 2012), camelid *KLRH* sequences represent only remnants of a full gene sequence. Although mRNA for bovine KLRJ was described (Storset et al., 2003), the precise splicing of camelid *KLRJ* and possible expression as a functional receptor remains to be verified. The low polymorphism of camelid *KLRK* is comparable to the limited polymorphism of this gene in humans and mice encoding an activating receptor NKG2D for diverse ligands (reviewed in Lanier, 2015).

The genomic organization of the NKC is similar in both domestic camel species and in *V. pacos*. The functionally important polymorphism of NKC genes is limited, with one monomorphic gene and six genes with two to three allelic protein variants in *C. dromedarius*. An even higher number of monomorphic *KLR* genes (three functional and two potential pseudogenes) was observed in *C. bactrianus*, and its *KLRC2* seems to be a pseudogene. It is not clear how this low NKC variation can be related to the fact that no "HLA-E-like" molecule has been found to date outside of simians and rodents and to our recent finding that the MHC gene cluster containing *HLA-E* in humans has been lost in camels, similarly to cattle and pigs (Plasil et al., 2019). Interestingly, *V. pacos* seems to be more polymorphic in NKC, at both the genomic and protein levels, despite the limited number of individual genomes analyzed.

Within the LRC region, no *KIR* genes have expanded, while *LILR* genes expanded both activating and inhibitory family members. As for the NKC, the same overall organization of the LRC with *FCAR*, *NCR1*, *KIR*, and *LILRB1* genes, three *LILRB* genes encoding a 4-Ig-like domain receptor, low variable *KIR3DL* and *LILRB1*, and unresolved *LILR* gene fragments was observed in *C. bactrianus*. The polymorphism of *KIR3DL* was similar in the dromedary (seven possible protein variants) and the alpaca (seven functional protein variants), while the alpaca seems to be more variable in the *LILRB1* gene (seven vs. three protein variants, respectively). Unfortunately, we were unable to analyze further alpaca *LILR* genes, mainly due to a lack of correct full-length gene reference sequences and/or to a low coverage of NGS reads available in public databases.

NCR1, NCR2, and NCR3 are the major activating receptors on human NK cells (reviewed in Koch et al., 2013). The *NCR1* gene is located within the LRC; *NCR2* and *3* are located on the human chromosome 6, with *NCR3* mapping within the MHC (Lanier, 2005). The chromosome location of *NCR1*, *NCR2*, and *NCR3* genes in the dromedary corresponds to the human homologues, suggesting an orthologous nature of the *NCR* sequences retrieved. However, only *NCR1* and *NCR2* have a structure of functional genes, while *NCR3* appears to be a pseudogene. The *NCR1*, *NCR2*, and *NCR3* genes of *C. bactrianus* and *V. pacos* are very similar in terms of their genomic locations, sequence homologies, and genomic variation.

We are aware of the limitations due to the quality of the current assembly; however, the clusters of the *LILR* sequences identified in the phylogenetic trees indicated, similarly to NKC genes, the individuality of each of the genes. Although the purpose of this study was to outline the general organization of the two NKR complexes in terms of major gene families represented, and their location within NKC and LRC, further work is needed to definitively resolve the complex structure of LRC region, and a detailed characterization of individual *LILR* genes and pseudogenes is needed.

Taken together, this first report on NKR genes in camelids revealed features characteristic for NKC and LRC of *Tylopoda*. Despite close phylogenetic relationships to cattle, important differences in the NKC and LRC genomic organization and their polymorphism were observed. On the other hand, many similarities with pigs were found. The data presented here increase our genetic knowledge of the innate immune variation in dromedaries and Bactrian camels and contribute to studies of NKR evolution in mammals. The results of this project add to our understanding of specificities of the camel immune system and its functions and represent a prerequisite for future investigations on MHC/NKR interactions in health and disease.

DATA AVAILABILITY STATEMENT

The camel datasets generated for this study can be found in the NCBI´s GenBank®. The sequences obtained for NKC genes have accession numbers MK644262 - MK644413. The LRC gene sequences for both camel species have accession numbers MK644414 - MK644532. The NCRs sequences have accesion numbers MK473784 - MK473840.

ETHICS STATEMENT

The dromedary DNA samples in this study were transferred from previous projects [Austrian Science Fund (FWF) P1084-B17 and P24706-B25; PI: PB] and originated either from plucked hair, EDTA blood collected commensally on Whatman FTA cards (Sigma-Aldrich, Vienna, Austria) during routine veterinary controls or from DNA extracts sent by collaborators under bilateral agreements. Samples were imported with permits from the Austrian Ministry of Labour, Social Affairs, Health and Consumer Protection. All Bactrian camel samples were collected commensally during veterinary procedures for previous research projects (GACR 523/09/1972; PI: PH). All samples were collected by a licensed veterinarian in compliance with all ethical and professional standards.

AUTHOR CONTRIBUTIONS

JF made NKC annotation, designed primers, carried out PCR for NKR genes, and analyzed data. JO made all NGS mappings. AJ made LRC annotation and carried out PCR for *LILRB1*. JE provided CamDro2 whole genome sequence and annotation. JW made microsatellite definition and analysis. AK designed the microsatellite project. PB designed the project. PH designed the

project and the NKR study. JF and PH drafted the manuscript. JO, AJ, and JW wrote paragraphs. JE and PB edited the manuscript. All authors read, commented on, and approved the final version of the manuscript.

FUNDING

This work was supported by the Austrian Science Fund FWF project P29623-B25, by CEITEC-Central European Institute of Technology with research infrastructure supported by the project CZ.1.05/1.1.00/02.0068 financed from European Regional Development Fund and by the Czech National Sustainability Programme NPU LQ1601. The microsatellite study was supported by Internal Grant Agency of the Faculty of Agronomy, Mendel University, in Brno (IGA FA MENDELU No. IP 057/2017).

ACKNOWLEDGMENTS

Dr. Martin Plášil (CEITEC-VFU, Brno) is greatly acknowledged for his technical assistance in DNA isolation and preparation of NGS libraries as well as carrying out of all MiSeq™ sequencing runs. We thank all camel owners and veterinarian colleagues for their kind cooperation during sample collection, O. Abdelhadi, A. Abdussamad, H. Burgsteiner, B. Faye, G. Gassner, J. Juhasz, G. Konuspayeva, D. Modry, P. Nagy, A. and J. Perret, M. Qablan, R. Saleh, and M. Sloboda.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00620/full#supplementary-material

REFERENCES

Abdallah, H. R., and Faye, B. (2013). Typology of camel farming system in Saudi Arabia. *Emir. J. Food Agric.* 25, 250–260. doi: 10.9755/ejfa.v25i4.15491

Allan, A. J., Sanderson, N. D., Gubbins, S., Ellis, S. A., and Hammond, J. A. (2015). Cattle NK cell heterogeneity and the influence of MHC class I. *J. Immunol.* 195, 2199–2206. doi: 10.4049/jimmunol.1500227

Antonacci, R., Mineccia, M., Lefranc, M., Ashmaoui, H. M. E., Lanave, C., Piccinni, B., et al. (2011). Expression and genomic analyses of Camelus dromedarius T cell receptor delta (TRD) genes reveal a variable domain repertoire enlargement due to CDR3 diversification and somatic mutation. *Mol. Immunol.* 48, 1384–1396. doi: 10.1016/j.molimm.2011.03.011

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Braud, V. M., Allan, D. S., O'Callaghan, C. A., Söderström, K., D'Andrea, A., Ogg, G. S., et al. (1998). HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* 391, 795–799. doi: 10.1038/35869

Breese, M. R., and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* 29, 494–496. doi: 10.1093/bioinformatics/bts731

Carrillo-Bustamante, P., Kesmir, C., and de Boer, R. J. (2016). The evolution of natural killer cell receptors. *Immunogenetics* 68, 3–18. doi: 10.1007/s00251-015-0869-7

Ciccarese, S., Vaccarelli, G., Lefranc, M., Tasco, G., Consiglio, A., Casadio, R., et al. (2014). Characteristics of the somatic hypermutation in the Camelus dromedarius T cell receptor gamma (TRG) and delta (TRD) variable domains. *Dev. Comp. Immunol.* 46, 300–313. doi: 10.1016/j.dci.2014.05.001

Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* 63, 41–49. doi: 10.1016/j.ymeth.2013.06.027

Daws, M. R., Ke-Zheng, D., Zinöcker, S., Naper, Ch., Kveberg, L., Hedrich, H. J., et al. (2012). Identification of an MHC class I ligand for the single member of a killer cell lectin-like receptor family, KLRH1. *J. Immunol.* 189, 5178–5184. doi: 10.4049/jimmunol.1201983

De Meyer, T., Muyldermans, S., and Depicker, A. (2014). Nanobody-based products as research and diagnostic tools. *Trends Biotechnol.* 32, 263–270. doi: 10.1016/j.tibtech.2014.03.001

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Dimasi, N., and Biassoni, R. (2005). Structural and functional aspects of the Ly49 natural killer cell receptors. *Immunol. Cell. Biol.* 83, 1–8. doi: 10.1111/j.1440-1711.2005.01301.x

Elbers, J. P., Rogers, M. F., Perelman, P. L., Proskuryakova, A. A., Serdyukova, N. A., Johnson, W. E., et al. (2019). Improving Illumina assemblies with Hi-C and long reads: an example with the North African dromedary. *Mol. Ecol. Resour.* 00, 1–12. doi: 10.1111/1755-0998.13020

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2016). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16, 314–324. doi: 10.1111/1755-0998.12443

Gossner, C., Danielson, N., Gervelmeyer, A., Berthe, F., Faye, B., Aaslav, K. K., et al. (2014). Human-dromedary camel interactions and the risk of acquiring zoonotic Middle East respiratory syndrome coronavirus infection. *Zoonoses Public Health.* 63, 1–9. doi: 10.1111/zph.12171

Guethlein, L. A., Norman, P. J., Hilton, H. G., and Parham, P. (2015). Co-evolution of MHC class I and variable NK cell receptors in placental mammals. *Immunol. Rev.* 267, 259–282. doi: 10.1111/imr.12326

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41, 95–98.

Hamerman, J. A., Ogasawara, K., and Lanier, L. L. (2005). NK cells in innate immunity. *Curr. Opin. Immunol.* 17, 29–35. doi: 10.1016/j.col.2004.11.001

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, E., Songa, E. B., et al. (1993). Naturally occurring antibodies devoid of light chains. *Nature* 363, 446–448. doi: 10.1038/363446a0

Hao, L., Klein, J., and Nei, M. (2006). Heterogenous but conserved natural killer receptor gene complexes in four major orders of mammals. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3192–3197. doi: 10.1073/pnas.0511280103

Hemida, M. G., Elmoslemany, A., Al-Hizab, F., Alnaeem, A., Almathen, F., Faye, B., et al. (2017). Dromedary camels and the transmission of Middle East respiratory syndrome coronavirus (MERS-CoV). *Transbound. Emerg. Dis.* 64, 344–353. doi: 10.1111/tbed.12401

Hogan, L., Bhuju, S., Jones, D. C., Laing, K., Trowsdale, J., Butcher, P., et al. (2012). Characterisation of bovine leukocyte Ig-like receptors. *PLoS One* 7, e34291. doi: 10.1371/journal.pone.0034291

Hussen, J., Shawaf, T., Al-herz, A. I., Alturaifi, H. R., and Alluwaimi, A. M. (2017). Reactivity of commercially available monoclonal antibodies to human CD antigens with peripheral blood leucocytes of dromedary camels (Camelus dromedarius). *Open Vet. J.* 7, 150–156. doi: 10.4314/ovj.v7i2.12

Ito, M., Maruyama, T., Saito, N., Koganei, S., Yamamoto, K., and Matsumoto, N. (2006). Killer cell lectin-like receptor G1 binds three members of the classical cadherin family to inhibit NK cell cytotoxicity. *J. Exp. Med.* 203, 289–295. doi: 10.1084/jem20051986

Jirimutu, Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., et al. (2012). Genome sequences of wild and domestic Bactrian camels. *Nat. Commun.* 3, 1202. doi: 10.1038/ncomms2192

Kelley, J., Walter, L., and Trowsdale, J. (2005). Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet.* 2, e27. doi: 10.1371/journal.pgen.0010027

Koch, J., Steinle, A., Watzl, C., and Mandelboim, O. (2013). Activating natural cytotoxicity receptors of natural killer cells in cancer and infection. *Trends Immunol.* 34, 182–191. doi: 10.1016/j.it.2013.01.003

Lanier, L. L. (1998). NK cell receptors. *Annu. Rev. Immunol.* 16, 359–393. doi: 10.1146/annurev.immunol.16.1.359

Lanier, L. L. (2005). NK cell recognition. *Annu. Rev. Immunol.* 23, 225–274. doi: 10.1146/annurev.immunol.23.021704.115526

Lanier, L. L. (2015). NKG2D receptor and its ligands in host defense. *Cancer Immunol. Res.* 3, 575–582. doi: 10.1158/2326-6066.CIR-15-0098

Lanier, L. L., Corliss, B., Wu, J., and Phillips, J. H. (1998). Association of DAP12 with activating CD94/NKG2C NK cell receptors. *Immunity* 8, 693–701. doi: 10.1016/S1074-7613(00)80574-9

Li, H. (2013).Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*®. Accessed January 21, 2019

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAM tools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Marsch, S. G. E., Parham, P., Dupont, B., Geraghty, D. E., Trowsdale, J., Middleton, D., et al. (2003). Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics* 55, 220–226. doi: 10.1007/s00251-003-0571-z

Martin, A. M., Kulski, J. K., Witt, C., Pontarotti, P., and Christiansen, F. T. (2002). Leukocyte Ig-like receptor complex (LRC) in mice and men. *Trends Immunol.* 23, 81–88. doi: 10.1016/S1471-4906(01)02155-X

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Megersa, B., Markemann, A., Angassa, A., Ogutu, J. O., Piepho, H., and Zarate, A. V. (2014). Livestock diversification: an adaptive strategy to climate change and rangeland ecosystem changes in southern Ethiopia. *Hum. Ecol.* 42, 509–520. doi: 10.1007/s10745-014-9668-2

Mossad, A. A., Elbagoury, A. R., Khalid, A. M., Waters, W. R., Tibary, A., Hamilton, M. J., et al. (2006). Identification of monoclonal antibody reagents for use in the study of immune response in camel and water buffalo. *Proc. Int. Sci. Conf. Camels* 13, 2391–2411.

Muyldermans, S., Baral, T. N., Retamozzo, V. C., De Baetselier, P., De Genst, E., Kinne, J., et al. (2009). Camelid immunoglobulins and nanobody technology. *Vet. Immunol. Immunopathol.* 128, 178–183. doi: 10.1016/j.vetimm.2008.10.299

Parham, P., and Moffett, A. (2013). Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat. Rev. Immunol.* 13, 133–144. doi: 10.1038/nri3370

Plasil, M., Mohandesan, E., Fitak, R. R., Musilova, P., Kubickova, S., Burger, P. A., et al. (2016). The major histocompatibility complex in Old World camelids and low polymorphism of its class II genes. *BMC Genomics* 17, 167. doi: 10.1186/s12864-016-2500-1

Plasil, M., Wijkmark, S., Elbers, J. P., Oppelt, J., Burger, P. A., and Horin, P. (2019). The major histocompatibility complex of Old World camelids: class I and class I-related genes. *HLA* 93, 203–215. doi: 10.1111/tan.13510

Saether, P. C., Westgaard, I. H., Hoelsbrekken, S. E., Benjamin, J., Lanier, L. L., Fossum, S., et al. (2008). KLRE/I1 and KLRE/I2: a novel pair of heterodimeric receptors that inversely regulate NK cell cytotoxicity. *J. Immunol.* 181, 3177–3182. doi: 10.4049/jimmunol.181.5.3177

Sanderson, N. D., Norman, P. J., Guethlein, L. A., Ellis, S. A., Williams, C., Breen, M., et al. (2014). Definition of the cattle killer cell Ig-like receptor gene family: comparison with aurochs and human counterparts. *J. Immunol.* 193, 6016–6030. doi: 10.4049/jimmunol.1401980

Schenkel, A. R., Kingry, L. C., and Slayden, R. A. (2013). The Ly49 gene family. A brief guide to the nomenclature, genetics, and role in intracellular infection. *Front. Immunol.* 4, 90. doi: 10.3389/fimmu.2013.00090

Schwartz, J. C., Gibson, M. S., Heimeier, D., Koren, S., Phillippy, A. M., Bickhart, D. M., et al. (2017). The evolution of the natural killer complex; a comparison between mammals using new high-quality genome assemblies and targeted annotation. *Immunogenetics* 69, 255–269. doi: 10.1007/s00251-017-0973-y

Schwartz, J. C., and Hammond, J. A. (2018). The unique evolution of the pig LRC, a single KIR but expansion of LILR and a novel Ig receptor family. *Immunogenetics* 70, 661–669. doi: 10.1007/s00251-018-1067-1

Steeland, S., Vandenbroucke, R. E., and Libert, C. (2016). Nanobodies as therapeutics: big opportunities for small antibodies. *Drug Discov. Today* 21, 1076–1113. doi: 10.1016/j.drudis.2016.04.003

Stephens, M., and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73, 1162–1169. doi: 10.1086/379378

Storset, A. K., Slettedal, I. Ö., Williams, J. L., Law, A., and Dissen, E. (2003). Natural killer cell receptors in cattle: a bovine killer cell immunoglobulin-like receptor multigene family contains members with divergent signaling motifs. *Eur. J. Immunol.* 33, 980–990. doi: 10.1002/eji.200323710

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Mol. Biol. Evol.* 9, 678–687. doi: 10.1093/oxfordjournals.molbev.a040752

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. *Evolution* 28, 2731–2739. doi: 10.1093/molbev/msr121

Trowsdale, J., Barten, R., Haude, A., Stewart, C. A., Beck, S., and Wilson, M. J. (2001). The genomic context of natural killer receptor extended gene families. *Immunol. Rev.* 181, 20–38. doi: 10.1034/j.1600-065X.2001.1810102.x

Vaccarelli, G., Antonacci, R., Tasco, G., Yang, F., Giordano, L., El Ashmaoui, H. M., et al. (2012). Generation of diversity by somatic mutation in the Camelus dromedarius T-cell receptor gamma variable domains. *Eur. J. Immunol.* 42, 3416–3428. doi: 10.1002/eji.201142176

Vance, R. E., Kraft, J. R., Altman, J. D., Jensen, P. E., and Raulet, D. H. (1998). Mouse CD94/NKG2A is a natural killer cell receptor for the nonclassical major histocompatibility complex (MHC) class I molecule Qa-1(b). *J. Exp. Med.* 188, 1841–1848. doi: 10.1084/jem.188.10.1841

Vivier, E., Raulet, D. H., Moretta, A., Caligiuri, M. A., Zitvogel, L., Lanier, L. L., et al. (2011). Innate or adaptive immunity? The example of natural killer cells. *Science* 331, 44–49. doi: 10.1126/science.1198687

Watson, E. E., Kochore, H. H., and Dabasso, B. H. (2016). Camels and climate resilience: adaptation in northern Kenya. *Hum. Ecol.* 44, 701–713. doi: 10.1007/s10745-016-9858-1

Wernery, U., and Kinne, J. (2012). Foot and mouth disease and similar virus infections in camelids: a review. *Rev. Sci. Tech. Off. Int. Epiz.* 31, 907–918. doi: 10.20506/rst.31.3.2160

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5, 5188. doi: 10.1038/ncomms6188

Zidan, M., and Pabst, R. (2008). Unique microanatomy of ileal Peyer's patches of the one humped camel (Camelus dromedarius) is not age-dependent. *Anat. Rec.* 291, 1023–1028. doi: 10.1002/ar.20697

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Futas, Oppelt, Jelinek, Elbers, Wijacki, Knoll, Burger and Horin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 July 2019
doi: 10.3389/fgene.2019.00692

Genome-Wide Identification of Microsatellites and Transposable Elements in the Dromedary Camel Genome Using Whole-Genome Sequencing Data

Reza Khalkhali-Evrigh[1], Nemat Hedayat-Evrigh[2], Seyed Hasan Hafezian[1], Ayoub Farhadi[1] and Mohammad Reza Bakhtiarizadeh[3]*

[1] Department of Animal Breeding and Genetics, Sari Agricultural Sciences and Natural Resources University, Sari, Iran, [2] Department of Animal Science, University of Mohaghegh Ardabili, Ardabil, Iran, [3]Department of Animal and Poultry Science, College of Aburaihan, University of Tehran, Tehran, Iran

Transposable elements (TEs) along with simple sequence repeats (SSRs) are prevalent in eukaryotic genome, especially in mammals. Repetitive sequences form approximately one-third of the camelid genomes, so study on this part of genome can be helpful in providing deeper information from the genome and its evolutionary path. Here, in order to improve our understanding regarding the camel genome architecture, the whole genome of the two dromedaries (Yazdi and Trodi camels) was sequenced. Totally, 92- and 84.3-Gb sequence data were obtained and assembled to 137,772 and 149,997 contigs with a N50 length of 54,626 and 54,031 bp in Yazdi and Trodi camels, respectively. Results showed that 30.58% of Yazdi camel genome and 30.50% of Trodi camel genome were covered by TEs. Contrary to the observed results in the genomes of cattle, sheep, horse, and pig, no endogenous retrovirus-K (ERVK) elements were found in the camel genome. Distribution pattern of DNA transposons in the genomes of dromedary, Bactrian, and cattle was similar in contrast with LINE, SINE, and long terminal repeat (LTR) families. Elements like RTE-BovB belonging to LINEs family in cattle and sheep genomes are dramatically higher than genome of dromedary. However, LINE1 (L1) and LINE2 (L2) elements cover higher percentage of LINE family in dromedary genome compared to genome of cattle. Also, 540,133 and 539,409 microsatellites were identified from the assembled contigs of Yazdi and Trodi dromedary camels, respectively. In both samples, di-(393,196) and tri-(65,313) nucleotide repeats contributed to about 42.5% of the microsatellites. The findings of the present study revealed that non-repetitive content of mammalian genomes is approximately similar. Results showed that 9.1 Mb (0.47% of whole assembled genome) of Iranian dromedary's genome length is made up of SSRs. Annotation of repetitive content of Iranian dromedary camel genome revealed that 9,068 and 11,544 genes contain different types of TEs and SSRs, respectively. SSR markers identified in the present study can be used as a valuable resource for genetic diversity investigations and marker-assisted selection (MAS) in camel-breeding programs.

Keywords: *Camelus dromedarius*, *de novo* assembly, repetitive sequence, breeding strategies, next-generation sequencing

OPEN ACCESS

Edited by:
Elena Ciani,
University of Bari Aldo Moro, Italy

Reviewed by:
René Massimiliano Marsano,
University of Bari Aldo Moro, Italy
Pablo Orozco-terWengel,
Cardiff University, United Kingdom

****Correspondence:***
Nemat Hedayat-Evrigh
nhedayat@uma.ac.ir

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *04 January 2019*
Accepted: *02 July 2019*
Published: *26 July 2019*

Citation:
Khalkhali-Evrigh R, Hedayat-Evrigh N, Hafezian SH, Farhadi A and Bakhtiarizadeh MR (2019) Genome-Wide Identification of Microsatellites and Transposable Elements in the Dromedary Camel Genome Using Whole-Genome Sequencing Data. Front. Genet. 10:692. doi: 10.3389/fgene.2019.00692

INTRODUCTION

Approximately, 42 to 46 million years ago (Mya), ancestors of extant camels appeared in the North America (Honey at al., 1998). The divergence between Camelini (Old World camels) and Lamini (New World camels) occurred in the Early Miocene (Rybczynski et al., 2013). Migration of Old World and New World camel's ancestor into Eurasia (*via* Bering Isthmus) and South America (*via* Isthmus of Panama), respectively (Heintzman et al., 2015), placed them in two different evolutionary paths. Estimated time for splitting of Old World camels into dromedary and Bactrian camels is approximately 4.4 Mya (Wu et al., 2014). Throughout the past, 4.9 to 7.2 million years (Camelini migration time to Eurasia), Old World camels have become adapted to the deserts of Asia and Africa (Wu et al., 2014) and are used as pack animal for nomads and low-income rural populations (Khalkhali-Evrigh et al., 2018).

Iran is mostly covered by arid or semi-arid regions. Some areas face with increasing population pressure, shortage of water resources, and risk of desertification (Heshmati and Squires, 2013). The evolution has donated some skills to camels enabling them to survive and reproduce in the mentioned condition. Climatic changes and traditional religious and cultural values of Iranians created a high potential for camel breeding in Iran. Currently, dromedary camels are considered as an important supplier of protein for people living in the desert areas in Iran.

It is well known that eukaryotic genome contains a large fraction of repetitive DNA, mainly tandem repeats (satellites, minisatellites, and microsatellites) and interspersed elements [or transposable elements (TEs) (Biscotti et al., 2015). These elements are important components of genomes and are responsible for genome size differences seen across eukaryotes (Sotero-Caio et al., 2017). The sequence, frequency, organization, structure, and location of the repeated units are mainly specific to each species (Pezer et al., 2012).

Since Barbara McClintock discovered TEs (mobile genetic elements) in mid-1940s, many studies have been carried out to understand their genomic function. Generally, TEs are scattered throughout the mammalian genome; however, they are in higher abundance in the heterochromatin (Pardue et al., 1996) and are classified into two classes. Class I (retrotransposon) includes long terminal repeat (LTR) retrotransposon and non-LTR retrotransposon that use an RNA as intermediate to jump (Lerat, 2010), while class II (DNA transposon) uses DNA as intermediate and can be divided into three subclasses including elements that use cut-and-paste, rolling-circle replication (Helitron element), and self-replicating mechanisms (Maverick/Polinton element) for their transposition (Feschotte and Pritham, 2007). Studies revealed that TEs play influential role in regulation of some mammalian gene expression (Medstrand et al., 2005) as well as a source of genetic innovation (Brandt et al., 2005), and also, they contribute in genomes restructuring to enhance the host's ability to respond to stress (Kidwell and Lisch, 2001).

Microsatellites or simple sequence repeats (SSRs) are known as one of the most variable types of DNA sequences in the genome of many species (Ellegren, 2004). SSRs are randomly distributed across the genome of most eukaryotes; also, they are codominant and highly polymorphic (Abdul-Muneer, 2014). Therefore, microsatellites are informative and widely used for the analysis of genetic diversity within and between populations as well as for evolutionary investigation among the livestock species (Hampton et al., 2004).

Next-generation sequencing technology along with recent advances in assembly strategies made it possible to walk along the genomes and achieve better understanding of them. First whole-genome sequencing of dromedary camel and alpaca (Wu et al., 2014) and Bactrian camels (Jirimutu et al., 2012) has opened a new window for researchers regarding the genome analysis of camelid. Contrary to domesticated species such as cattle, sheep, horse, chicken, and pig, researchers have paid less attention to camels, especially on Iranian camel breeds. Improvement of camel breeding status and designing appropriate breeding schemes require extensive knowledge about camels. Researchers in the light of analysis of camel genome from different aspects can achieve a deep understanding of this species. In this study, we sequenced and assembled (*de novo*) genomes of two Iranian dromedary camels. The distribution and frequency of different TEs and microsatellites were further characterized throughout their genomes. Also, the findings of this study were compared with findings related to other mammalian genomes.

MATERIALS AND METHODS

DNA Extraction and Sequencing

In this study, genomes of two Iranian dromedary camels were sequenced belonging to Yazd station in Yazd Province (YaD) and Trod station in Semnan Province (TrD). Blood samples were taken from Jugular vein and were stored in −20°C till use. DNA extraction was performed using RBC Mini Kit for mammalian blood (Real Biotech Corporation, RBC, South Korea). The extracted DNA was quantified using a NanoDrop, and the identified 260/280 ratio was equal to 1.90 and 1.80 for YaD and TrD, respectively; then quality of the DNA samples was assessed using gel electrophoresis in 1% agarose gel. A library was generated with an average insert size of ~360 bp, and two lanes of 100 bp paired-end sequencing were carried out using Illumina HiSeq 2000 system (Illumina, San Diego, CA) for each camel.

Quality Filtering and *de Novo* Assembly

Firstly, FastQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control of raw sequencing reads. Quality check showed no adapter contamination in raw sequencing reads. Quality filtering of short reads was performed using maximum information (MAXINFO) approach of the Trimmomatic v0.36 (Bolger et al., 2014) with a target length of 40 and strictness value of 0.5. Also, reads with a length less than 40 bp were discarded in the final step of quality filtration. Second, quality control was carried out after quality filtering, and results obtained from FastQC in this step indicated that quality of reads has increased, especially in 3' end of reads (**Supplementary Figure 1**).

We used CLC Genomics Workbench v9.0 (CLC Bio, Aarhus, Denmark) to perform the *de novo* assembly of trimmed reads using the following parameters: three for mismatch cost, three

for insertion and deletion cost, 0.5 for length fraction, and 0.8 for similarity fraction. In order to test the completeness of assembled genomes of our samples and Targui breed (accession number: GCA_001640815.1), BUSCO strategy was applied using *Mammalia* (containing 4,104 genes) and *Vertebrata* (containing 2,586 genes) datasets. In fact, BUSCO uses sets of Benchmarking Universal Single-Copy Orthologs from OrthoDB (www.orthodb.org) to report completeness degree of assemblies (Simao et al., 2015). Also, we used MUMmer3 (nucmer script) for comparison of our genomes with existing dromedary reference (accession number: GCA_000803125.1); results are presented in the **Supplementary File 1** (YaD with reference) and file2 (TrD with reference).

TEs Identification

Homology-based approach was used to identify TEs in assembled genomes of Iranian dromedaries. RepeatMasker v4.0.7 program (http://www.repeatmasker.org) was used to search for known TEs using the combination of Repbase v20170127 and Dfam databases. Repbase is a comprehensive database of repetitive sequences from human and other eukaryotic organisms. We used RMBlast v2.6.0 as a sequence search engine for RepeatMasker with Smith-Waterman cutoff 255 (based on Fitak et al., 2016). It is worth mentioning that RMBlast is a RepeatMasker compatible version of the standard NCBI BLASTn program, which is used to search for repeats. Also, *de novo* identification of TEs in genomes of Iranian dromedaries was performed using RepeatModeler v1.0.11 program (http://www.repeatmasker.org/RepeatModeler). In fact, RepeatModeler assists in automating the runs of RECON (Bao and Eddy, 2002) and RepeatScout (Price et al., 2005) as two *de novo* repeat finding programs, to analyze our genomes.

SSRs Identification

Assembled genomes were searched for identifying the microsatellites using MIcroSAtellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/) with motif size ranging from mono-nucleotide to octo-nucleotide. The minimum repeat numbers were defined as 12 for mono-, 6 for di-, 5 for tri- and tetra, 4 for penta- and hexa-, and 3 for hepta- and octo-nucleotide repeat SSRs. Microsatellite motifs that interrupted by 100 nucleotides were considered as compound microsatellites. Also, several mammalian assembled genomes were downloaded and searched for microsatellite loci, including Arabian dromedary camel, Bactrian camel, alpaca, horse, cattle, sheep, and human (downloaded from NCBI Reference Sequence Database, RefSeq). Our goal was to produce a dataset to compare with the results obtained for Iranian dromedaries. In order to extract comparable results, discovery of SSRs was done for all genomes with same parameters. Accession numbers of downloaded assembled genomes are presented in **Supplementary Table 1**.

Annotation of Repeats

We employed MAKER v2.31.10 pipeline (Cantarel et al., 2008; Holt and Yandell, 2011) to annotation of Iranian dromedary genome (YaD). It is a powerful tool for annotation of newly sequenced genome and also updating existing annotations. Protein and mRNA sequences for dromedary and Bactrian camels were obtained from NCBI as input for MAKER to homology-based gene identification. MAKER aligns these sequences to newly assembled genome for discovery of putative genes. CD-HIT (CD-HIT-2d; Li and Godzik, 2006) was used to prepare a database from dromedary and Bactrian protein sequences, as input for MAKER. In fact, CD-HIT-2d compares two protein datasets and one of the similar sequences in each cluster at a certain threshold (95% identity in our work) is reported as output. Finally, BEDtools v2.25.0 (Quinlan, 2014) was performed to find genes containing the TEs and SSRs. Based on the 80–80–80 rule (Wicker et al., 2007), genes containing TEs shorter than 80 bp were filtered out. It should be mentioned that DAVID v.6.8 (Huang et al., 2008) was used for functional enrichment analysis of genes containing some class of TEs (all subfamilies of MIRs) and SSRs. The calculated p-values were corrected using the Benjamini correction for multiple testing, and enriched terms were considered statistically significant at p-adjusted < 0.1.

RESULTS

De Novo Assembly and Completeness Assessment of Iranian Dromedary Genome

Sequencing of the samples using Illumina HiSeq 2000 platform in paired end yielded a total of 920,366,954 (92 Gb) and 843,455,144 (84.3 Gb) raw reads for YaD and TrD, respectively. Filtering to a threshold of Q20 sequence quality produced 899,714,102 and 826,229,484 clean reads, which were *de novo* assembled using CLC Genomics Workbench assembler. The clean reads were applied to assemble the genome for each camel, separately. The assembly process yielded 137,772 and 149,997 contigs with total consensus genome size of 1.94 Gb in YaD and TrD. The contig N50 length was equal to 54.6 and 54 kb for YaD and TrD, respectively, indicating good quality assembly for further downstream analysis (**Table 1**). The averaged GC contents of the assembled genome were 41.52% and 41.58% in YaD and TrD, respectively, which was slightly higher than the reported GC content for African dromedary (41.3%; Fitak et al., 2016), Arabian dromedary (41.2%), alpaca (41.4%; Wu et al., 2014), and wild Bactrian camel (41.28%; Jirimutu et al., 2012). The results of completeness test on assembled genomes using BUSCO revealed that 93.7% of the 4,104 genes in the *Mammalia* dataset were present in YaD (74.1% complete genes, 19.6% fragmented genes) and TrD (73.9% complete genes, 19.8%

TABLE 1 | Summary of the YaD and TrD genome assembly.

Contigs	YaD	TrD
N25 (bp)	97,128	95,611
N50 (bp)	54,626	54,031
N75 (bp)	26,694	26,620
Longest contig	466,683	604,268
Average contig length (bp)	14,101	12,980
Counts of contigs	137,772	149,997
Total bases (Gb)	1.94	1.94

fragmented genes) genomes. Also, *Vertebrata* dataset was used to investigate the completeness of assemblies, and it was found that 94.9% and 94.8% of vertebrate genes were present in the YaD and TrD assemblies, respectively.

Identification of Repeats in Iranian Dromedary Genomes

The amount of repetitions of the assembled genomes was assessed in order to annotate and determine how their components are organized, including TE and SSR diversity, distribution, and dynamics. To do this, a *de novo* and homology-based approach was applied to identify microsatellites and TEs using the assembled genomes, respectively. Based on RepeatMasker outputs, totally, 30.58% of Yazdi and 30.5% of Trodi camel genomes were composed of TEs. However, due to high similarity of repetitive sequences in two Iranian dromedary camels as well as similar assembled genome sizes of both, average values of different repetitive elements were used for discussion (**Table 2**). In fact, results of homology-based method on repeats identification showed that the total length of TEs content was equal to 594.1 Mb (30.54% of assembled genome) in YaD and TrD camels. Also, results of *de novo* based identification of TEs for YaD and TrD, as additional information for future studies, are presented in **Supplementary Table 2**.

Results of TE annotation revealed that 635 types of them were located in 9,068 genes (**Supplementary Table 3A**). LINEs with 50,653 copies had the most distribution in the genic regions followed by SINEs, LTRs, and DNA transposons with 28,701, 16,839, and 11,442 copies, respectively. Some TEs such as Eulors and UCONs (173 elements) were grouped as unclassified elements. About 28,444 of identified genic SINEs belonged to all subfamilies of MIRs (MIR, MIRb, MIRc, MIR3, and MIR1_Amn) that have influenced 6,920 genes.

Profiles of SSRs

SSRs are extremely useful molecular markers for study on genomic diversity among individuals as well as among populations and different species (Saeed et al., 2016). Therefore, genome-wide identification of these makers can be considered as a valuable genomic resource for population characterization. We investigated the SSR distribution on Iranian dromedaries as well as on the seven mammalian genomes. Frequencies of different types of microsatellite in dromedary camels and other mammals are presented in **Figure 1** and **Table 3**. The highest and lowest percentages of SSRs in whole-genome content belong to human (0.79% of assembled genome) and horse (0.32% of assembled genome), respectively.

The results of scanning the assembled genomes revealed the presence of 86,121 and 69,061 contigs possessing microsatellite motifs in YaD and TrD, respectively. Totally, 540,133 and 539,409 microsatellites loci were found, corresponding to 9.1 Mb of repeat bases which represents 0.47% of the total bases of the genomes in both Iranian dromedary genomes. Among these, 50,805 and 50,896 sequences belonged to compound types in YaD and TrD, respectively. The frequency of microsatellite repeats was found to range from one motif per 2,435.1 bp in human to one motif per 5,147.9 bp in horse genome. For YaD and TrD, the density of SSRs was equal to one motif per 3,596.8 and 3,609.4 bp, respectively, indicating that camel genome has a medium SSR density compared to the investigated mammalian genomes.

Results of SSR annotation showed that there are 42,636 SSRs in the 11,544 genes (**Supplementary Table 3B**). Tri- and hexa-nucleotide SSRs include ~10% of all genic SSRs. As expected, the most SSRs in genic regions belonged to mono- (21,512 motifs) followed by di-nucleotide (12,831 motifs) SSRs. Results of gene ontology (GO) analysis for 1,000 genes with the largest number of all types of SSRs revealed that these genes are present in places such as membrane and cell cortex. Also, we found that terms such as ATP binding, ATPase activity, ligase activity, etc. were significantly enriched in molecular function category (**Supplementary Table 4A**).

TABLE 2 | Summary of identified transposable elements for Iranian dromedaries, African dromedary, and Bactrian camel.

TEs	Iranian dromedary			African dromedary		Bactrian camel	
	Numbers	Length (bp)	%	Numbers	%	Numbers	%
SINEs	**458,654**	**68,422,832**	**3.51**	**473,387**	**3.43**	**560,273**	**3.92**
Alu/B1	0	0	0.00	7	0.00	0	0.00
MIRs	452,075	67,605,477	3.48	463,927	3.38	460,330	3.38
LINEs	**838,990**	**354,299,228**	**18.22**	**1,009,426**	**19.28**	**883,074**	**19.32**
LINE1	483,215	260,189,224	13.38	642,633	14.57	552,674	14.82
LINE2	301,928	81,743,253	4.20	312,130	4.10	280,625	3.92
L3/CR1	39,626	8,761,454	0.45	40,821	0.44	38,504	0.44
RTE	13,046	3,412,114	0.18	–	–	10,913	0.14
LTRs	**286,303**	**103,155,112**	**5.31**	**324,636**	**5.43**	**321,595**	**5.81**
ERVL	82,052	35,068,299	1.80	80,984	1.72	76,625	1.63
ERVL-MaLRs	140,115	47,792,407	2.45	138,020	2.35	133,362	2.31
ERV-classI	39,386	13,849,285	0.71	81,938	1.07	77,025	1.63
ERV-classII	0	0	0.00	571	0.00	23,095	0.10
DNA elements	**331,140**	**68,223,263**	**3.50**	**341,448**	**3.44**	**282,696**	**3.00**
hAT-Charlie	188,154	36,658,424	1.88	186,819	1.79	175,700	1.68
TcMar-Tigger	53,998	14,327,829	0.74	66,902	0.81	44,496	0.68
Total	**2,581,776**	**625,581,334**	**30.54**	**2,905,840**	**31.58**	**2,628,996**	**32.05**
Reference		Present study		Fitak et al. (2016)		Jirimutu et al. (2012)	

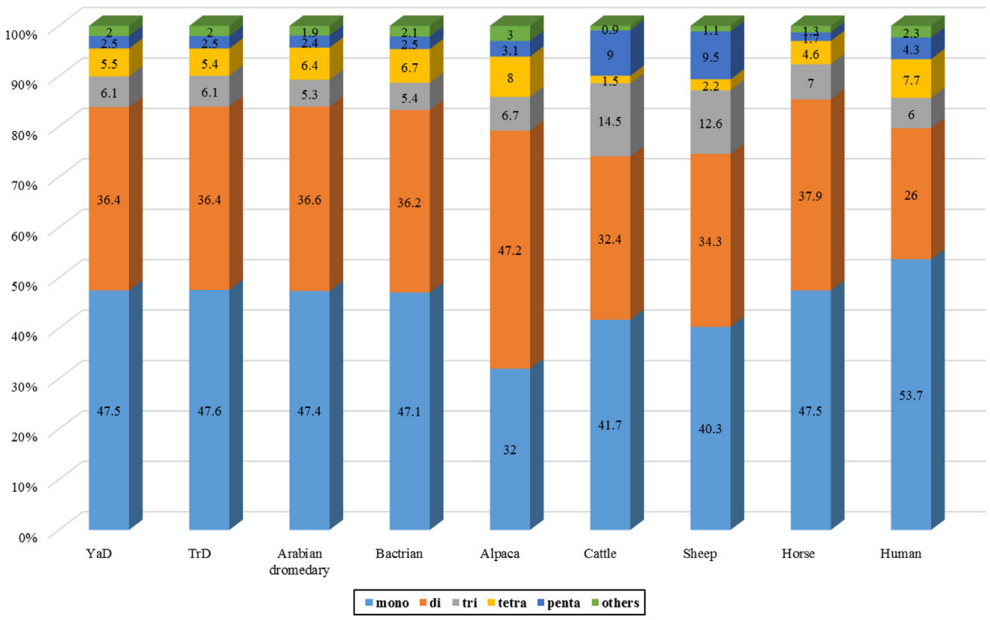


FIGURE 1 | Frequency of different SSR motifs across YaD, TrD, and other seven mammalian genomes.

TABLE 3 | Distribution of different classes of SSRs in YaD, TrD, and other seven mammalian genomes.

Species	Mono	Di	Tri	Tetra	Penta	Hexa	Hepta	octo	Total (bp)
YaD	3,670,743	3,429,832	638,283	781,068	329,445	111,366	97,839	59,368	9,117,944
TrD	3,669,900	3,417,918	636,564	778,964	325,930	110,706	97,790	58,208	9,095,980
Arabian dromedary	4,121,293	3,859,778	685,755	1,390,196	401,970	129,516	97,986	89,456	10,775,950
Bactrian camel	4,082,129	3,791,158	615,105	1,147,212	366,725	125,718	99,785	92,632	10,320,464
Alpaca	2,065,967	4,102,030	627,525	1,100,084	348,420	123,546	152,278	72,032	8,591,882
Cattle	5,045,498	4,858,676	2,046,501	254,296	1,537,845	54,444	94,374	26,576	13,918,210
Sheep	4,657,293	4,889,182	1,770,183	392,980	1,589,625	62,922	78,764	57,312	13,498,261
Horse	3,344,142	3,025,388	577,437	532,720	180,225	47,934	64,001	37,848	7,809,695
Human	12,287,665	6,754,988	1,537,530	3,001,280	1,439,420	291,282	304,525	149,072	25,765,762

DISCUSSION

As a prerequisite to genomic scale studies as well as development of breeding programs in Iranian dromedaries, we used the first whole-genome resequencing data of individual camels from two distinct geographical regions (Khalkhali-Evrigh et al., 2018) for mining of repetitive content. Contig N50 lengths obtained in present study were found to be longer than African dromedary (40.2 kb; Fitak et al., 2016) and Targui breed dromedary (31.5 kb; GenBank accession: GCA_001640815.1) but shorter than Arabian dromedary (69.1 kb; Wu et al., 2014) and wild Bactrian camel (90.3 kb; Jirimutu et al., 2012). The results of completeness test related to previously published assembled genome in contig level (BioProject: PRJNA310822; Targui breed) revealed that, despite the various libraries in mentioned project, better completeness was achieved in this study compared to Targui breed (**Supplementary Table 5**). The shorter assembled genomes (1.94 Gb) in present study, compared to 2.01 Gb in Arabian dromedary (Wu et al., 2014) and 2.08 Gb in Targui dromedary, may be attributed to the lack of libraries with long insert size in the present study. However, results of BUSCO implied that our genome assembly can be comparable to previously reported assemblies.

Identification of the amount of repetitions regions in genomes can be used in refining genome assembling and annotation. Furthermore, proper annotation of repeats provides information on the evolutionary mechanisms involved in species differentiation and how they diversified over the evolutionary process (Mehrotra and Goyal, 2014). Results of RepeatMasker for TEs investigation in Iranian dromedary genome were found to be close to African breed dromedary (31.58%; Fitak et al., 2016), Bactrian camel (30.37%), and alpaca (32.14%; Wu et al., 2014), but less than horse (47.3%; Adelson et al., 2010) and cattle genomes (46.5%; Adelson et al., 2009). This number of repetitive regions in these genomes can be attributed to C-value paradox, which is the observed lack of correlation between increases in DNA content and an organism's complexity. In other words, there is a correlation between genome size in eukaryotes and repetitive regions (not gene content)

(Eddy, 2012). Therefore, higher repetitive content of horse, cattle, and human genomes may be due to their bigger genome size.

Previous findings showed that lengths of non-repetitive regions of mammalian genomes are similar, as cattle (genome size of 2.67 Gb), sheep (genome size of 2.61 Gb), horse (genome size of 2.47 Gb), dromedary camel (genome size of 2.05 Gb), and Bactrian camel (genome size of 1.99 Gb) have 1.37, 1.41, 1.25, 1.32, and 1.36 Gb unique genomic content, respectively. Here, non-repetitive content of assembled genomes was 1.32 Gb for both Iranian dromedary genomes that are in agreement with other mammalian genomes. Seemingly, calculated non-repetitive regions in mammalian genomes are conserved section of their genomes; however, proof of this claim requires further studies.

In the present study, TEs were found to contribute in 30.54% of assembled genome of Iranian dromedary camels. It is reported that LINEs and SINEs are very old TEs in mammalian genomes (about thousands of millions of years ago) (Medstrand et al., 2005). L1 is mostly located in AT-rich regions and is dominant LINE in mammalians, while LINE2 (L2) is uniformly distributed throughout genome (Gu et al., 2000). Also, abundance of LINEs within genes is less than their abundance in upstream and downstream regions of genes. On the other hand, SINEs are overrepresented within genes in comparison with LINEs (Medstrand et al., 2005). In the present study, we found that LINE elements were the most prevalent interspersed repeats and contributed 354.3 Mb (18.22%) of the total assembled genomes, which was slightly lower than African dromedary (19.28%) and Bactrian camel (19.32%). Additionally, in LINE class, LINE1 (L1) and RTE elements with 483,215 (13.38%) and 13,046 (0.18%) copies were the most and least frequent elements, respectively. It is well known that LINE RTE (BovB) elements are considered as the gift of squamates to ruminants. Presumably, horizontal transfer of BovB elements has been done by ticks (especially *Amblyomma* and *Bothriocroton* species) as common vectors between squamates, ruminants, monoterms, and African mammals (Walsh et al., 2013). BovB elements comprised 10.70% and 11.70% of cattle (Adelson et al., 2009) and sheep genome, respectively, while this value for dromedary camel, Bactrian camel, horse, and pig were 0.035%, 0.051%, 0.079%, and 0.034%, respectively. Further studies are needed to understand why distribution pattern of BovB elements is different throughout domesticated mammalian genomes.

Content of SINEs was less than LINEs and comprised only 3.51% of the total genome length. It is well known that Alu elements are mostly enriched in GC-rich or gene-rich regions, and they are considered as abundant and conserved repeat family in primate genomes (Gu et al., 2000). In this study, no Alu elements were found in the Iranian dromedary camel genomes. This finding was in agreement with Bactrian camels; however, Fitak et al. (2016) reported seven Alu sequences in African dromedary camel. Due to the lack of mentioned seven Alu annotations in African dromedary camel genome, we were unable to perform specific similarity searches. Therefore, we cannot determine if the Iranian dromedary camel's genome actually lacks Alus or if this is a false-negative result due to the sequence divergence of Alu elements of camels and primates.

Mammalian-wide interspersed repeats (MIRs) are another important member of SINE family. Positive association has been found between existence of one TE in genic regions and tissue-specific gene expression for MIRs (Jjingo et al., 2011). Jjingo et al. (2014) studied on human genome and revealed that MIRs are rich source of transcription factor binding sites compared to random genomic regions. Therefore, enrichment of MIRs within enhancers influences gene expression level as well as tissue-specific gene expression. We found that 67.6 Mbp (3.48%) of Iranian dromedary camel genome was covered by MIRs, which was slightly higher than Bactrian camel and African camel (3.38%). Study on MIR elements in RNA-seq level can help us to better understand the roles of MIRs in camel genome. Among all the TE elements identified in Iranian dromedary camel genome, 286,303 copies were classified as LTR elements including ERVL, ERVL-MaLRs, ERV-classI, and ERV-classII (**Table 2**).

The content of different interspersed repetitive sequence families including SINEs, LINEs, LTRs, and DNA transposons in dromedary, Bactrian, and cattle genomes is shown in **Figure 2**. A distinct pattern was observed in the proportion of each subfamily elements of SINEs, LINEs, and LTRs for cattle genome compared to camelid genomes. Mammalian LTR transposon (MaLR) was the most frequent member of LTRs in all three genomes. MaLR elements [with 1.5–10-kb length (Gu et al., 2000)] were inserted in mammalian genome about 70 million years ago (Bènit et al., 1999). The previous findings about TEs in horse genome showed that most of MaLR elements in genic regions were located in coding region (96 elements) in comparison with 3'-UTR (6 elements) and 5'-UTR (0 element) (Ahn et al., 2013). Comparison of DNA elements in dromedary, Bactrian, and cattle genome showed that hAT-charlie element is the most abundant DNA elements among all three genomes.

It has been shown that 0.05% and 0.07% of cattle and horse genomes are composed of ERVK (belonging to LTRs class) elements, respectively (Adelson et al., 2009; Adelson et al., 2010). However, ERVK element was absent in dromedary's genome that is in accordance with African dromedary (Fitak et al., 2016) and Bactrian camel (Jirimutu et al., 2012). ERVKs are one of the youngest members of endogenous retroviruses (ERVs) family (Katoh and Kurata, 2013), and the conservation of the specific protein binding site in some ERVKs probably reflects the regulatory role of them for the nearby located genes in the human genome (Akopov et al., 1998). Unlike the LINE, SINE, and LTR families, distribution of DNA transposon elements was similar among dromedary, Bactrian, and cattle.

In this study, we found that 3.5% of dromedary genomes belong to DNA transposons, in which hAT-charlie and TcMar-Tigger elements were the most abundant members of this family. Percentages of DNA elements in the genomes of human, mouse, cattle (Adelson et al., 2009), horse (Adelson et al., 2010), and Bactrian camel (Jirimutu et al., 2012) are 3, 0.89, 1.96, 3.1, and 3, respectively. Scientists believed that the last activity of DNA transposons in mammalian genomes has occurred at least 40 million years ago. However, Ray et al. (2008) provided evidences for recent activity of hAT and Helitron elements in bat (*Myotis lucifugus*) lineage. These results pave the way for further investigation of active elements in mammalian genomes as well as their effects.

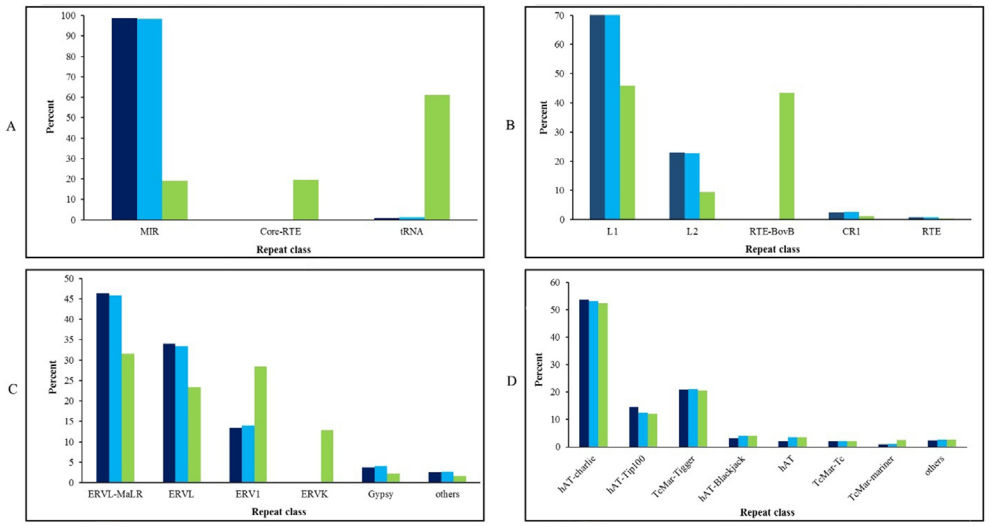


FIGURE 2 | The content of SINE **(A)**, LINE **(B)**, LTR **(C)**, and DNA transposon elements **(D)** in the genomes of Iranian dromedaries (dark blue), Bactrian camel (blue), and cattle (green).

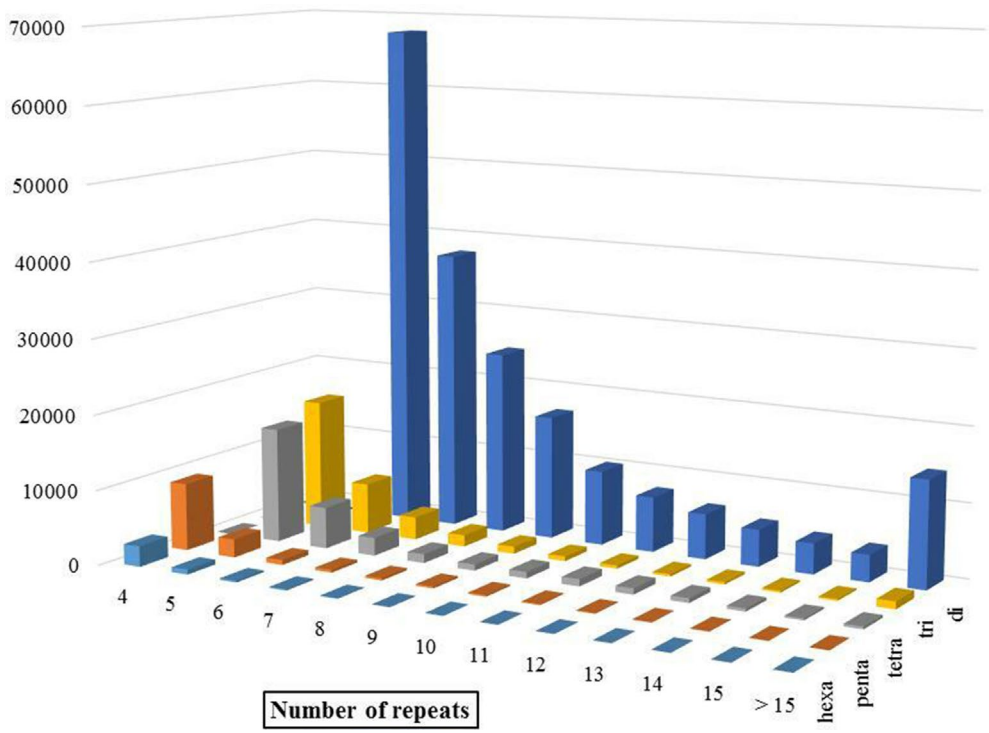


FIGURE 3 | Distribution of different SSR motifs in Iranian dromedary genome.

Discovery and mining of genomic repeats such as SSRs are considered as a successful approach in genetic analysis, linkage mapping, and marker-assisted breeding (Pandey et al., 2017). The results of assessment of microsatellite distribution in whole genome of Iranian dromedaries revealed that the number of microsatellites decreased with the increase in size. Also, for each class of repeats, the number of motifs decreased as the number of repeats increased (**Figure 3**). For example, di-nucleotide motif with six repeats contains 34.4% (67,724) of total di-nucleotides, while this value for 15 repeats was equal to 1.8% (3,610). Number of the identified microsatellite motifs ranged from 469,380 (alpaca) to 1,336,255 (human). SSR contents in YaD and TrD were lower than those of Arabian dromedary (10.7 Mb) and Bactrian camel (10.3 Mb). One possible explanation for the discrepancy among SSR contents

is different length and assembly levels of the genomes; as in this study, the genomes were assembled at contig levels, but Arabian and Bactrian camel genomes were made at the scaffold level. Unexpectedly, pattern of microsatellites in alpaca, as a camelid, was different from dromedary and Bactrian camels. Mono-nucleotide (47.1% to 47.6%) was the most frequent motif in Old World camels, whereas di-nucleotide motifs (47.2%) were dominant in alpaca. The results revealed that lowest and highest mono-nucleotide motifs belonged to alpaca (32%) and human (53.7%), respectively. The highest content of tetra-nucleotide motifs was assigned to alpaca followed by human, Old World camels, horse, sheep, and cattle.

(T)n motifs were found to be the most abundant repeat in YaD, TrD, Bactrian camel, cattle, sheep, horse, and human, whereas (A)n was the most abundant motif for Arabian dromedary and alpaca. Among di-nucleotide SSRs, AC/GT type was enriched in the genomes followed by AT/TA and AG/CT types in all under study mammalians. In case of tri-nucleotide SSRs, around 26–28% of them belonged to AAT/ATT in camelid, whereas shares of AAT/ATT motifs in horse and human genome were 25.76% and 39.91%, respectively. The highest percentage of tri-nucleotide motifs in sheep and cattle genome belonged to AGC/CTG motif with a share of about 68.3% and 75.51% of all tri-nucleotide motifs, respectively. The AAAC/GTTT motif was the most frequent tetra-nucleotide motif in dromedaries and Bactrian camel, while the AAAT/ATTT motif had the highest number of tetra-nucleotide in others. In the case of hexa-nucleotide, AAAAAC/GTTTTT motif was the most replicated motif in all genomes except cattle genome (**Supplementary File 3**). Assessment of most frequent motifs in different classes of SSRs revealed that there is a high similarity between mammalians in this regard. This similarity along with non-random distribution of SSRs throughout genome and functional roles of them (Vieira et al., 2016) may reflect the importance of these repeated sequences in evolution process in mammalians.

Results of genome-wide identification of microsatellites in Iranian dromedaries revealed that AT-rich motifs were dominant in all classes of repeats. In this context, A/T motifs in mono-nucleotide, AC/GT in di-nucleotide, AAC/GTT in tri-nucleotide, AAAC/GTTT in tetra-nucleotide, AAAAC/GTTTT in penta-nucleotide, AAAAAC/GTTTTT in hexa-nucleotide, AAAAAAC/GTTTTTT in hepta-nucleotide, and AAAAAAAC/GTTTTTTT in octo-nucleotide were found as the most abundant motif type.

Furthermore, 10 abundant microsatellite motifs with highest frequencies for each genome were studied. Totally, it was found that more than 74% of all microsatellites in each genome belonged to top 10 frequent motifs (**Table 4**). Also, a very similar pattern of these microsatellites was observed among dromedaries, Bactrian camel, alpaca, horse, and even human. In mentioned species, all of 10 motifs were composed of mono- and di-nucleotide.

Then, unique SSR loci were considered and human was found as the most diverse species with 5,544 SSR types. For alpaca, the number of SSR motif types was equal to 5,296 followed by TrD (5,122), YaD (5,112), Bactrian camel (4,995), Arabian dromedary (4,853), horse (3,869), sheep (3,597), and cattle (3,223). Also, 1,629 SSR motifs were identified in common between camelid genomes. Moreover, 847 motif types were identified, which were shared among all evaluated genomes in this study (**Supplementary Table 6**). Species-specific motifs in camelid genomes were 782 (17 in tetra-, 101 in penta-, 195 in hexa-, 296 in hepta-, and 173 in octo-nucleotide). On the other hand, 209 motifs (5 tetra-, 65 penta-, 25 hexa-, 77 hepta-, and 37 octo-nucleotide) were found in common among human, cattle, sheep, and horse, whereas they were not observed in camelid genomes (**Supplementary Table 7**).

Generally, the findings of the present study showed that the content and distribution of the identified repetitive regions were similar (not the same) among Iranian dromedaries and the other camel breeds with sequenced genome. The differences in the repetitive content of these breeds can be attributed to different factors, such as different evolutionary origin or discrepancy in the assembly stage of these genome projects. Of note, repetitive sequences are much harder to assemble in a *de novo* manner and tend to form smaller contigs, resulting in different content of repetitive distribution in closely related breeds.

Because of the potential regulatory role of MIRs in mammalian genomes (Wang et al., 2015), GO analysis was applied on genes containing this class of TEs. For this, 1,000 genes were selected with the most number of any subfamilies of MIRs for classification based on biological process, cellular component, and molecular function. For these genes, we found no significantly enriched GO term in biological process and

TABLE 4 | 10 SSRs with most frequency in YaD, TrD, and other seven mammalian genomes.

Rank	YaD	TrD	Arabian dromedary	Bactrian camel	Alpaca	Cattle	Sheep	Horse	Human
1	T	T	A	T	A	T	T	T	T
2	A	A	T	A	T	A	A	A	A
3	AC	AC	AC	AC	AC	TG	TG	TG	AC
4	TG	TG	TG	TG	TG	AC	AC	AC	TG
5	AT	AT	AT	AT	AT	AT	AT	TA	AT
6	TA	TA	TA	TA	TA	AGC	TA	CA	TA
7	GT	GT	GT	GT	GT	TA	CA	AT	GT
8	CA	CA	CA	CA	CA	CA	GT	GT	CA
9	TC	TC	TC	TC	TC	GT	AGC	TC	TC
10	AG	AG	AG	AG	AG	TGC	ACTGA	AG	AG
From all (%)	78.81	78.83	79.06	79.02	74.96	76.15	74.27	78.54	77.29

cellular component but two for molecular function including ATP binding and calcium ion binding (**Supplementary Table 4B**). Optimum energy metabolism is vital for camels due to the low food and harsh living environment of them; therefore, the presence of these elements in the genes related to energy metabolism in camel genome because of regulatory roles and possible link between MIRs and enhancers (Jjingo et al., 2014) is considered an interesting result. Absolutely, more and deeper studies are needed to prove the relevance of MIRs with important genes (like genes associated with energy metabolism) in the camel genome.

Here, for the first time, DNA sequencing technology was used for *de novo* genome assembly of Iranian dromedaries. Although the amount of applied sequencing data was not adequate for whole-genome assembly, it could help us to obtain an overview regarding the repetitive elements in the genome. Furthermore, in order to generate a comprehensive annotation of the Iranian dromedary camels' repeatome, we used a computational approach. The results revealed that, on average, 594.1 and 9.1 Mb of Iranian dromedary's genome length are made up of TEs and SSRs, respectively.

The finding of this study will be applied as a valuable resource for further studies on camel breeding, especially on Iranian dromedary's breeds. The large number of camel's SSR markers developed in this study established a valuable resource for investigation of genetic diversity, marker-assisted selection (MAS) and may improve the development of breeding programs in Iranian dromedary camels in the future. This study was like shedding light on a part of the camel genome. Absolutely, conduction of more studies would provide more information and awareness about genomic features of camels. Increasing genome-wide information about camels could improve the designed strategies used for its maintenance and breeding.

ETHICS STATEMENT

All animal care and experiments were approved by the animal science committee of the University of Mohaghegh Ardabili, Iran. Also, all experiments were performed in accordance with a routine guideline, which is acceptable by this committee. It is worth to mention that, for reducing stress of animals, positive rewards such as petting were implemented in the conditioning to regular handling prior to restraint for blood collection.

AUTHOR CONTRIBUTIONS

RK-E and MRB developed the idea and analyzed the data. NH-E collected the samples and obtained funding and required equipment for the project. SHH supervised the project, AF advised the project, and RK-E, NH-E, and MRB interpreted the results and wrote the manuscript.

ACKNOWLEDGMENTS

This work supported by the University of Mohaghegh Ardabili. Also, the authors thank Yazd and Trod stations for the collaboration on collecting blood samples from the camels.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00692/full#supplementary-material. Also, some extra supplementary files including outputs of RepeatMasker, RepeatModeler, MISA and annotation of repeats, are archived in the Dryad Digital Repository under doi: 10.5061/dryad.10h185k.

REFERENCES

Abdul-Muneer, P. M. (2014). Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genet. Res. Int*. 2014, 691759 doi: 10.1155/2014/691759

Adelson, D., Raison, J., Garber, M., and Edgar, R. (2010). Interspersed repeats in the horse (Equus caballus); spatial correlations highlight conserved chromosomal domains. *Anim. Genet.* 41, 91–99. doi: 10.1111/j.1365-2052.2010.02115.x

Adelson, D. L., Raison, J. M., and Edgar, R. C. (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *PNAS* 106, 12855–12860. doi: 10.1073/pnas.0901282106

Ahn, K., Bae, J. H., Gim, J. A., Lee, J. R., Jung, Y. D., Park, K. D., et al. (2013). Identification and characterization of transposable elements inserted into the coding sequences of horse genes. *Genes Genom.* 35, 483–489. doi: 10.1007/s13258-013-0057-9

Akopov, S. B., Nikolaev, L. G., Khil, P. P., Lebedev, Y. B., and Sverdlov, E. D. (1998). Long terminal repeats of human endogenous retrovirus K family (HERV-K) specifically bind host cell nuclear proteins. *FEBS Lett.* 421, 229–233. doi: 10.1016/S0014-5793(97)01569-X

Bao, Z., and Eddy, S. R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276. doi: 10.1101/gr.88502

Bénit, L., Lallemand, J. B., Casella, J. F., Philippe, H., and Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J. Virol.* 73, 3301–3308.

Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. P. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Res.* 23, 415–420. doi: 10.1007/s10577-015-9499-z

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Brandt, J., Schrauth, S., Veith, A. M., Froschauer, A., Haneke, T., Schultheis, C., et al. (2005). Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* 345, 101–111. doi: 10.1016/j.gene.2004.11.022

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi: 10.1101/gr.6743907

Eddy, S. R. (2012). The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* 22, 898–899. doi: 10.1016/j.cub.2012.10.002

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435. doi: 10.1038/nrg1348

Feschotte, C., and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368. doi: 10.1146/annurev.genet.40.110405.090448

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2016). The *de novo* genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16, 314–324. doi: 10.1111/1755-0998.12443

Gu, Z., Wang, H., Nekrutenko, A., and Li, W. H. (2000). Densities, length proportions, and other distributional features of repetitive sequences in the human genome

estimated from 430 megabases of genomic sequence. *Gene* 259, 81–88. doi: 10.1016/S0378-1119(00)00434-0

Hampton, J. O., Spencer, P., Alpers, D. L., Twigg, L. E., Woolnough, A. P., Doust, J., et al. (2004). Molecular techniques, wildlife management and the importance of genetic population structure and dispersal: a case study with feral pigs. *J. Appl. Ecol.* 41, 735–743. doi: 10.1111/j.0021-8901.2004.00936.x

Heintzman, P. D., Zazula, G. D., Cahill, J. A., Reyes, A. V., MacPhee, R. D., and Shapiro, B. (2015). Genomic data from extinct North American Camelops revise camel evolutionary history. *Mol. Biol. Evol.* 32, 2433–2440. doi: 10.1093/molbev/msv128

Heshmati, G. A., and Squires, V. (2013). *Combating desertification in Asia, Africa and the Middle East*. New York: Springer. doi: 10.1007/978-94-007-6652-5

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491. doi: 10.1186/1471-2105-12-491

Honey, J. G., Harrison, J. A., Prothero, D. R., and Stevens, M. S. (1998). *Evolution of tertiary mammals of North America, volume 1: terrestrial carnivores, ungulates, and ungulatelike mammals*. Cambridge: Cambridge University Press.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44. doi: 10.1038/nprot.2008.211

Jirimutu, Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., et al. (2012). Genome sequences of wild and domestic Bactrian camels. *Nat. Commun.* 3, 1202. doi: 10.1038/ncomms2192

Jjingo, D., Huda, A., Gundapuneni, M., Mariño-Ramírez, L., and Jordan, I. K. (2011). Effect of the transposable element environment of human genes on gene length and expression. *Genome Biol. Evol.* 3, 259–271. doi: 10.1093/gbe/evr015

Jjingo, D., Conley, A. B., Wang, J., Mariño-Ramírez, L., Lunyak, V. V., and Jordan, I. K. (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA* 5, 14. doi: 10.1186/1759-8753-5-14

Katoh, I., and Kurata, S. I. (2013). Association of endogenous retroviruses and long terminal repeats with human disorders. *Front. Oncol.* 3, 234. doi: 10.3389/fonc.2013.00234

Khalkhali-Evrigh, R., Hafezian, S. H., Hedayat-Evrigh, N., Farhadi, A., and Bakhtiarizadeh, M. R. (2018). Genetic variants analysis of three dromedary camels using whole genome sequencing data. *PLoS One.* 13, e0204028. doi: 10.1371/journal.pone.0204028

Kidwell, M. G., and Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55, 1–24. doi: 10.1111/j.0014-3820.2001.tb01268.x

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* 104, 520–533. doi: 10.1038/hdy.2009.165

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Medstrand, P., Van de Lagemaat, L., Dunn, C. A., Landry, J. R., Svenback, D., and Mager, D.L. (2005). Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet. Genome Res.* 110, 342–352. doi: 10.1159/000084966

Mehrotra, S., and Goyal, V. (2014). Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics Proteomics Bioinformatics* 12, 164–171. doi: 10.1016/j.gpb.2014.07.003

Pandey, M., Kumar, R., Srivastava, P., Agarwal, S., Srivastava, S., Nagpure, N. S., et al. (2017). WGSSAT: a high-throughput computational pipeline for mining and annotation of SSR markers from whole ggenomes. *J. Hered.* 109, 339–343. doi: 10.1093/jhered/esx075

Pardue, M. L., Danilevskaya, O. N., Lowenhaupt, K., Slot, F., and Traverse, K. L. (1996). Drosophila telomeres: new views on chromosome evolution. *Trends Genet.* 12, 48–52. doi: 10.1016/0168-9525(96)81399-0

Pezer, Ž., Brajković, J., Feliciello, I., and Ugarković, Đ., (2012). Satellite DNA-mediated effects on genome regulation, in *Repetitive DNA* (Basel: Karger Publishers). doi: 10.1159/000337116

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics* 21, 351–358. doi: 10.1093/bioinformatics/bti1018

Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 47, 11–12. doi: 10.1002/0471250953.bi1112s47

Ray, D. A., Feschotte, C., Pagan, H. J., Smith, J. D., Pritham, E. J., Arensburger, P., et al. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res.* 18, 717–728. doi: 10.1101/gr.071886.107

Rybczynski, N., Gosse, J. C., Harington, C. R., Wogelius, R. A., Hidy, A. J., and Buckley, M. (2013). Mid-Pliocene warm-period deposits in the high Arctic yield insight into camel evolution. *Nat. Commun.* 4, 1550. doi: 10.1038/ncomms2516

Saeed, A. F., Wang, R., and Wang, S. (2016). Microsatellites in pursuit of microbial genome evolution. *Front. Microbiol.* 6, 1462. doi: 10.3389/fmicb.2015.01462

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Sotero-Caio, C. G., Platt, R. N., Suh, A., and Ray, D. A. (2017). Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* 9, 161–177. doi: 10.1093/gbe/evw264

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. d. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39, 312–328. doi: 10.1590/1678-4685-GMB-2016-0027

Walsh, A. M., Kortschak, R. D., Gardner, M. G., Bertozzi, T., and Adelson, D. L. (2013). Widespread horizontal transfer of retrotransposons. *PNAS* 110, 1012–1016. doi: 10.1073/pnas.1205856110

Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., et al. (2015). MIR retrotransposon sequences provide insulators to the human genome. *PNAS* 112, 4428–4437. doi: 10.1073/pnas.1507253112

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982. doi: 10.1038/nrg2165

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5, 5188. doi: 10.1038/ncomms6188

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer POtW declared their involvement as co-editors in the Research Topic, and confirm the absence of any other collaboration.

Copyright © 2019 Khalkhali-Evrigh, Hedayat-Evrigh, Hafezian, Farhadi and Bakhtiarizadeh. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 August 2019
doi: 10.3389/fgene.2019.00748

Casein Gene Cluster in Camelids: Comparative Genome Analysis and New Findings on Haplotype Variability and Physical Mapping

Alfredo Pauciullo[1], El Tahir Shuiep[2], Moses Danlami Ogah[3], Gianfranco Cosenza[4], Liliana Di Stasio[1] and Georg Erhardt[5]*

[1] Department of Agricultural, Forest and Food Sciences, University of Torino, Grugliasco, Italy, [2] Institute of Molecular Biology, University of Nyala, Nyala, Sudan, [3] Department of Animal Science, Nasarawa State University, Keffi, Shabu-Lafia, Nigeria, [4] Department of Agriculture, University of Napoli Federico II, Portici Italy, [5] Department for Animal Breeding and Genetics, Justus Liebig University, Gießen, Germany

OPEN ACCESS

Edited by:
Pamela Burger, University of Veterinary Medicine, Austria

Reviewed by:
Gaukhar Konuspayeva, Al-Farabi Kazakh National University, Kazakhstan
Gustavo Augusto Gutierrez Reynoso, National Agrarian University, Peru

***Correspondence:**
Alfredo Pauciullo
alfredo.pauciullo@unito.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *25 October 2018*
Accepted: *17 July 2019*
Published: *29 August 2019*

Citation:
Pauciullo A, Shuiep ET, Ogah MD, Cosenza G, Di Stasio L and Erhardt G (2019) Casein Gene Cluster in Camelids: Comparative Genome Analysis and New Findings on Haplotype Variability and Physical Mapping. Front. Genet. 10:748. doi: 10.3389/fgene.2019.00748

The structure of casein genes has been fully understood in llamas, whereas in other camelids, this information is still incomplete. In fact, structure and polymorphisms have been identified in three (*CSN1S1*, αs1-CN; *CSN2*, β-CN; *CSN3*, κ-CN) out of four casein genes, whereas controversial information is available for the *CSN1S2* (αs2-CN) in terms of structure and genetic diversity. Data from the genome analysis, whose assembly is available for feral camel, Bactrian, dromedary, and alpaca, can contribute to a better knowledge. However, a majority of the scaffolds available in GenBank are still unplaced, and the comparative annotation is often inaccurate or lacking.Therefore, the aims of this study are 1) to perform a comparative genome analysis and synthesize the literature data on camelids casein cluster; 2) to analyze the casein variability in two dromedary populations (Sudanese and Nigerian) using polymorphisms at *CSN1S1* (c.150G > T), *CSN2* (g.2126A > G), and *CSN3* (g.1029T > C); and 3) to physically map the casein cluster in alpaca. Exon structures, gene and intergenic distances, large insertion/deletion events, SNPs, and microsatellites were annotated. In all camelids, the *CSN1S2* consists of 17 exons, confirming the structure of llama *CSN1S2* gene. The comparative analysis of the complete casein cluster (~190kb) shows 12,818 polymorphisms. The most polymorphic gene is the *CSN1S1* (99 SNPs in Bactrian *vs*. 248 in dromedary *vs*. 626 in alpaca). The less polymorphic is the *CSN3* in the Bactrian (22 SNPs) and alpaca (301 SNPs), whereas it is the *CSN1S2* in dromedary (79 SNPs). In the two investigated dromedary populations, the allele frequencies for the three markers are slightly different: the allele C at *CSN1S1* is very rare in Nigerian (0.054) and Sudanese dromedaries (0.094), whereas the frequency of the allele G at *CSN2* is almost inverted. Haplotype analysis evidenced GAC as the most frequent (0.288) and TGC as the rarest (0.005). The analysis of R-banding metaphases hybridized with specific probes mapped the casein genes on chromosome 2q21 in alpaca. These data deepen the information on the structure of the casein cluster in camelids and add knowledge on the cytogenetic map and haplotype variability.

Keywords: camels, casein, haplotype, caseins genes mapping, interspersed element, microsatelite

INTRODUCTION

Camelids are the only living animals naturally spread over three continents: Africa (dromedaries), Asia (dromedaries and both wild and domesticated Bactrians), and South America (llamas, alpacas, vicunas, and guanacos). Camelids are popular also in Australia (mainly dromedaries) and Europe (llamas and alpacas). However, these populations are not indigenous, but imported from middle of 1800 in Australia (McKnight, 1969), and more recently in Europe. Since their domestication, Old and New world camelids have been exploited as multi-purpose animals for transportation (as beasts of burden), food (as source of milk and meat), but also kept for their fiber (wool and hair), and finally for entertainment (as riding animals). Therefore, these animals are of major economic and cultural importance for nomadic societies of Africa and Asia, as well as for the rural populations of South America.

Although their potential to survive on marginal resources in harsh environment, camelids have not been exploited as an important food source, and in particular of milk. For instance, only 10% of the total milk produced in the rural regions is of camel origin (Faye and Konuspayeva, 2012). Conversely, in the countries of the Gulf, intensive dromedary camel milk production in high-scale modernized unit has been already realized (Faye et al., 2002) and genetic improvement programs for the milk production have been implemented (Nagy et al., 2012).

The daily milk production of dromedary camels is estimated to vary between 3 and 10 kg during a lactation period of 12–18 months (Farah et al., 2007), depending on breed, stage of lactation, feeding, and management conditions, with an average content of 2.9% and 3.1% of protein and fat, respectively (Konuspayeva et al., 2009; Al hay and Al Kanhal, 2010). Data on daily milk production in Bactrians are more variable, depending also by the amount sucked by the calf. On average, it varies between 0.25 and 20 kg per day, with 3.9% of proteins and 5.3% of fat on average (for a review, see Zhao et al., 2015). Conversely, much lower yields were recorded in llamas, whose production ranges between 16 and 413 ml/day during a lactation period reaching a maximum of 220 days (Morin and Rowan, 1995) with an average content of 4.2% and 4.7% of protein and fat, respectively (Riek and Gerken, 2006). In alpacas, milk yield was assessed in a range from 0.4 to 1.2 L/day (Leyva and Markas, 1991).

As for ruminants, the main constituent of camel milk proteins are caseins. Caseins are coded by single autosomal genes, in order *CSN1S1* (αs1-casein), *CSN2* (β-casein), *CSN1S2* (αs2-casein), and *CSN3* (κ-casein), organized as a cluster in a DNA stretch of about 250 kb mapped on chromosome 6 in cattle, sheep, and goat (Rijnkels, 2002). Caseins have been recognized as a powerful molecular model for evolutionary studies (Kawasaki et al., 2011), and their genetic characterization in less investigated species is a useful tool for a better understanding of phylogenetic relationships among domesticated mammalian species and breeds.

In dromedary camels, *CSN2* and *CSN3* genes have been fully characterized (Pauciullo et al., 2013a; Pauciullo et al., 2014), whereas a partial genomic DNA sequence for *CSN1S1* was reported by Shuiep et al. (2013). The casein gene cluster has been investigated also in llama at mRNA level (Pauciullo and Erhardt, 2015) and protein level (Saadaoui et al., 2014), whereas only partial information is known for alpaca (Erhardt et al., 2017).

In dromedary camels, genetic polymorphisms have been identified in three out of four casein genes. Kappeler et al. (1998) described the first two genetic variants (A and B) of *CSN1S1*, which differ for eight amino acids (EQAYFHLE), skipped in A variant as consequence of the alternative splicing of the exon 18 (Erhardt et al., 2016). The C variant was identified at protein level by isoelectrofocusing (IEF) and confirmed at DNA level as polymorphism at the exon 5 (c.150G > T) responsible for the amino acid replacement p.30Glu > Asp (Shuiep et al., 2013). Recently, another variant (D) has been identified by IEF (Erhardt et al., 2016). Apparently, the sequence coding for this variant does not differ from that of the A allele, apart from an insertion of 11 bp in the intron 17, which may affect the spliceosome machinery then generating the skipping of the exon 18 (Erhardt et al., 2016). Genetic variants have been described also for the *CSN2* and *CSN3*. The SNP g.2126A > G at *CSN2* and g.1029T > C at *CSN3* are particularly relevant for changing consensus sequences for transcription factors (TATA-box and HNF-1, respectively) (Pauciullo et al., 2013a; Pauciullo et al., 2014). Conversely, controversial information on exons' number is available for the *CSN1S2* gene, and no SNP has been reported so far for the αs2-casein, despite a series of alternative splicing variants have been recently described by Ryskaliyeva et al. (2019). However, in this respect, useful data may derive from the genome analysis, whose assembly is available on line for feral, Bactrian, and dromedary camel, as well as for alpaca. The complete sequence is made of about 2,000 Mbases each species, but the isolated genomic scaffolds available in GenBank are still unplaced, and their annotation is almost completely lacking (Avila et al., 2014a). This observation underlines the need to acquire more data to help the annotation of the camel genome. Furthermore, considering the tight association among the casein genes, the estimation of the relationship between casein variants and milk production traits can be improved by considering the casein haplotypes instead of single genes.

The karyotype structure of camelids (2n = 74) and their similarities have been elucidated (Bunch et al., 1985; Di Berardino et al., 2006). However, lack of information exists in the cytogenetic mapping of genes, being located only few hundreds (Avila et al., 2014b; Perelman et al., 2018) not including casein loci that are important for their link with favorable/undesirable characteristic of coat color fibers, as observed in other species (Grosz and MacNeil, 1999).

Therefore, aims of the of this study are 1) to propose a revised and detailed comparative analysis of the casein cluster in Bactrian, dromedary, and alpacas using the feral camel genome as reference and the annotation available for all casein transcripts in llama; 2) to analyze the casein cluster variability in two dromedary populations (Sudanese and Nigerian) using genetic markers at *CSN1S1, CSN2,* and *CSN3*; and 3) to physically map the casein genes in alpaca.

MATERIALS AND METHODS

In order to accomplish the aims of the study, a dual approach was used. A multiple bioinformatics analysis of the genomes was achieved to elucidate the cluster and gene organization, the level of genetic diversity (SNP and microsatellites), the variability in the Interspersed elements, and the type of regulatory elements of the gene expression. A laboratory approach was accomplished to genotype and establish haplotypes in the dromedaries and to map cytogenetically the genes in alpaca.

Genome Comparative Analysis

The contig 039344 available in EMBL with the acc. no. AGVR01039100.1 and isolated from the whole genome sequence of the feral camel (Wang et al., 2012) was used as reference to establish sizes, positions, and orientations of the genes belonging to casein cluster. Scaffolds 146 (NW_011517196), 313 (NW_011591251), and 223 (KN269544) belonging respectively to Bactrian, dromedary, and alpaca genomes were used in the comparative analysis to describe differences in the casein cluster and to detect inter-specific genetic diversity.

Homology searches, comparison among sequences, and multiple alignments were achieved using MEGA 4 software (Tamura et al., 2007), whereas repeat masking was performed by Censor software (Kohany et al., 2006). Microsatellites were found by BioPHP Microsatellite repeat finder (http://insilico.ehu.es/mini_tools/microsatellites/). The main putative transcription factor binding sites were searched by TFBIND software considering 85% as minimum binding score.

Computational analysis of spliceosome specific sites was achieved by FruitFly software (http://www.fruitfly.org/seq_tools/splice.html), whereas the protein secondary structure was predicted by Jpred 4 software (http://www.compbio.dundee.ac.uk/jpred/), and the impact on protein biological functions was assessed by PROVEAN (Protein Variation Effect Analyser) software (http://provean.jcvi.org/index.php).

Genepop software was used to estimate allele frequencies and to test for Hardy-Weinberg equilibrium ($\chi 2$ test). Casein haplotype frequencies were estimated by PHASE ver.2.1 (Li and Stephens, 2003).

Ethics Approval Statement

Samples collection from dromedary followed all institutional and specific national guidelines for the care and use of laboratory animals. In particular, protocols were approved by Research and Ethics Committees of the Nasarawa State University (approval no: NSU/REC/AGRO10) for Nigerian camels and authorized by Ministry of Animal Resources and Fisheries (no number is available) for Sudanese dromedaries.

The collection of alpaca samples was done according to the German Animal Welfare Act. On the basis of article 8 (7) 2a of this law, no notification of or approval by the Animal Protection Unit of the Regional Council of Giessen, Germany, was necessary for this study.

Camelus dromedarius DNA Samples

A total of 267 blood samples were collected from dromedaries in Sudan and Nigeria. Samples were considered as representative of both countries because they were collected in different regions. Those from Sudan came from five areas: El Shuak (El Gadarif State), West Omdurman (Khartoum State), El Obeid (North Kordofan State), Nyala (South Darfur State), and Tamboul (El Butana area). Those from Nigeria came from Kano and Sokoto areas (North and North-west regions, respectively).

In particular, 198 Sudanese she-camels belonging to different ecotypes including Shanbali, Kahli, Lahaoi, and Arabi dromedary camels were provided by University of Nyala (Nyala, South Darfur, Sudan) and collected between years 2011 and 2012, whereas 69 Nigerian autochthonous dromedary camels were provided by Nasarawa State University (Nigeria) and collected between years 2016 and 2017.

DNA was isolated from blood leucocytes with the procedure already described by Sambrook et al. (1989).

DNA concentration and $OD_{260/280}$ ratio were measured with the Nanodrop ND-1000 Spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, USA).

Genotyping at Dromedary Camel *CSN1S1, CSN2,* and *CSN3* by PCR-RFLP Methods

Genotyping was carried out at DNA level using the methods described by Shuiep et al. (2013) for the c.150G > T at *CSN1S1* (allele C), Pauciullo et al. (2014) for the g.2126A > G at *CSN2*, and Pauciullo et al. (2013a) for the g.1029T > C at *CSN3*. Primer sequences, the thermal amplification conditions, and the list of restriction enzymes are reported in **Table 1**. PCR amplification was carried out using Bio-Rad T100 thermocycler (Bio-Rad). The digestion products were analyzed directly by electrophoresis in 2.5% agarose gel in 1X TBE buffer and stained with ethidium bromide.

Karyotyping and Probe Preparation

Peripheral blood cell cultures from two German alpacas were treated for late incorporation of BrdU (15 mg/ml) to obtain R-banding preparations. Hoechst 33258 (30 mg/ml) was simultaneously added to BrdU 6 h before harvesting to enhance the R-banding patterns. The alpacas were karyotyped according to standard methods for RBA-banding techniques (Iannuzzi and Di Berardino, 2008). Chromosome identification followed the R-banded ideogram of *Vicugna pacos* (2n = 74) chromosomes (Di Berardino et al., 2006). The R-banding preparations were further used for FISH analysis.

The casein gene probes were prepared by PCR amplification and cloning of five DNA fragments spread over the casein genes (primers are provided in **Table 1**) according to the method described by Pauciullo et al. (2013b). Labeling was carried out by standard nick translation reactions (Roche, Germany) using biotin-16-dUTP (Roche) as modified nucleotide. The probes were then used for FISH analysis.

TABLE 1 | Sequences and annealing temperature of the primers used for the genotyping by PCR-RFLP assays **(A)** and for preparation of the FISH probes covering the casein genes cluster **(B)**. All primers were designed on wild feral camel genome sequence available in gene bank (EMBL acc. no. AGVR01039100.1), and multiple alignment confirmed 100% similarity in the other camelids.

SNP (A)	GenBank ID		Primers	Annealing temperature (°C)	Size (bp)	Genotyping
CSN1S1 c.150G > T	JF429138	Forward:	5'-TGAACCAGACAGCATAGAG-3'	58.5	930	*SmlI*
		Reverse:	5'-CTAAACTGAATGGGTGAAAC-3'			
CSN2 g.2126A > G	HG969421	Forward:	5'-GTTTCTCCATTACAGCATC-3'	60.0	659	*HphI*
		Reverse:	5'-TCAAATCTATACAGGCACTT-3'			
CSN3 g.1029T > C	HE863813	Forward:	5'-CACAAAGATGACTCTGCTATCG-3'	62.0	488	*AluI*
		Reverse:	5'-GCCCTCCACATATGTCTG-3'			
Probe (B)	**Gene**		**Primers**	**Annealing temperature (°C)**	**Size (bp)**	**Position**
1	*CSN1S1*	Forward	5'-GTACCCAGAAGTCTTTCAA-3'	59.5	913	Exon 3
		Reverse	5'-CACTGCTAACTCAAGAATCT-3'			Exon 5
2	*CSN2*	Forward	5'-TTCACTTCTTTTCCTCCAC-3'	62.3	2433	Exon 1
		Reverse	5'-CCATTGTATTTGTGCAATATTA-3'			Intron 1
3	*CSN2*	Forward	5'-GATGAACAGCAGGATAAAATC-3'	56.0	657	Exon 7
		Reverse	5'-ATCACTGATCTGAACTAT-3'			Intron 7
4	*CSN1S2*	Forward	5'-AGCTGTAAGGAACATAAAGG-3'	60.5	1493	Exon 7
		Reverse	5'-TGTGGGGACTTCAGCTG-3'			Exon 8
5	*CSN3*	Forward	5'-TGCAGAGGTGCAAAACCA -3'	61.5	1337	Exon 4
		Reverse	5'-GCTAGTCTGTGTTGGTAGTAA-3'			Exon 5

Fluorescent *In Situ* Hybridization (FISH)

RPBI-FISH was performed according to Pauciullo et al. (2013b) and Pauciullo et al. (2016) with minor modifications. Briefly, 500 ng of labeled DNA from each of the nick translation reactions were combined and mixed together with competitor DNA. The probes were precipitated in ethanol 100% and then reconstituted in 7 µl hybridization solution (50% formamide in 2X SSC + 10% dextran sulfate), denatured at 75°C for 10 min, and incubated at 37°C for 60 min for pre-hybridization.

Fixed R-banding metaphase plates were stained with Hoechst 33258 (25 µg/ml) for 10 min, then washed, mounted in 2X SSC (pH 7.0), and exposed to UV light for 30 min to reinforce the banding. The slides were then denatured for 3 min in a solution of 70% formamide in 2X SSC (pH 7.0) at 75°C.

The hybridization mixture was applied to the slides and incubated in a moist chamber at 37°C for 3 days. Detection was performed three times with 1:400 fluorescein isothiocyanate (FITC)-avidin (Vector Laboratories, CA, USA) and 1:200 anti-avidin antibody (Vector Laboratories, CA, USA). Finally, slides were mounted with antifade/propidium iodide (3 µg/ml) and observed at 100× magnification with a Leica DM5500 fluorescence microscope equipped with FITC and Texas Red (TXRD) specific filters and provided with a CytoVision MB 8 image-analysis system (Leica Microsystems, Wetzlar, Germany).

A total of 30 randomly selected metaphase cells were examined per each alpaca to ensure the reliability of the probe signals by FISH. The hybridization efficiency was calculated as follows: FISH efficiency (%) is equal to the number of cells with hybridization signals present at the 2q21 region of both chromosomes 2 divided by the number of cells examined (Pauciullo et al., 2013b).

RESULTS

Multiple Bioinformatics Analysis

Cluster Organization

The caseins of camelids are encoded by four genes tightly clustered in a DNA fragment of about 190 kb. The organization and the orientation of the genes are highly conserved compared to all species studied to date, although with large differences in sizes partially due to a diverse number and natures of the interspersed repeated elements [short interspersed elements (SINEs), long interspersed elements (LINEs), microRNA (miR), etc.], partially due to genome expansion events and a higher number of genes present (**Figure 1**).

The first two casein genes (*CSN1S1* and *CSN2*) are close up (6.6 kb) compared to higher intergenic distances of the other casein genes (**Table 2**). For instance, a large distance (85.7 kb) exists between the *CSN1S2* and *CSN3*. In this interval, the *ODAM* gene was found, whereas no other known genes were found in the intergenic intervals *CSN1S1-CSN2* and *CSN2-CSN1S2*.

The comparative analysis of the genome sequence of wild feral camel (EMBL acc. no. AGVR01039100.1) with the annotated casein genes in dromedaries (HG969421; HE863813) and cDNAs of the whole cluster in llama (EMBL acc. nos. LK999986; LK999992; LK999989; LK999995) allowed the complete exon identification in all camelids. The *CSN1S1* is made of 20 exons in dromedary and 21 exons in the other camelids, the *CSN2* consisted of nine exons, the *CSN1S2* is arranged in 17 exons, and the *CSN3* is organized in five exons (**Table 2**). Splice donor and acceptor consensus sequences conforming to the 5'-GT/3'-AG rule were identified at the exon/intron boundaries. The average GC content

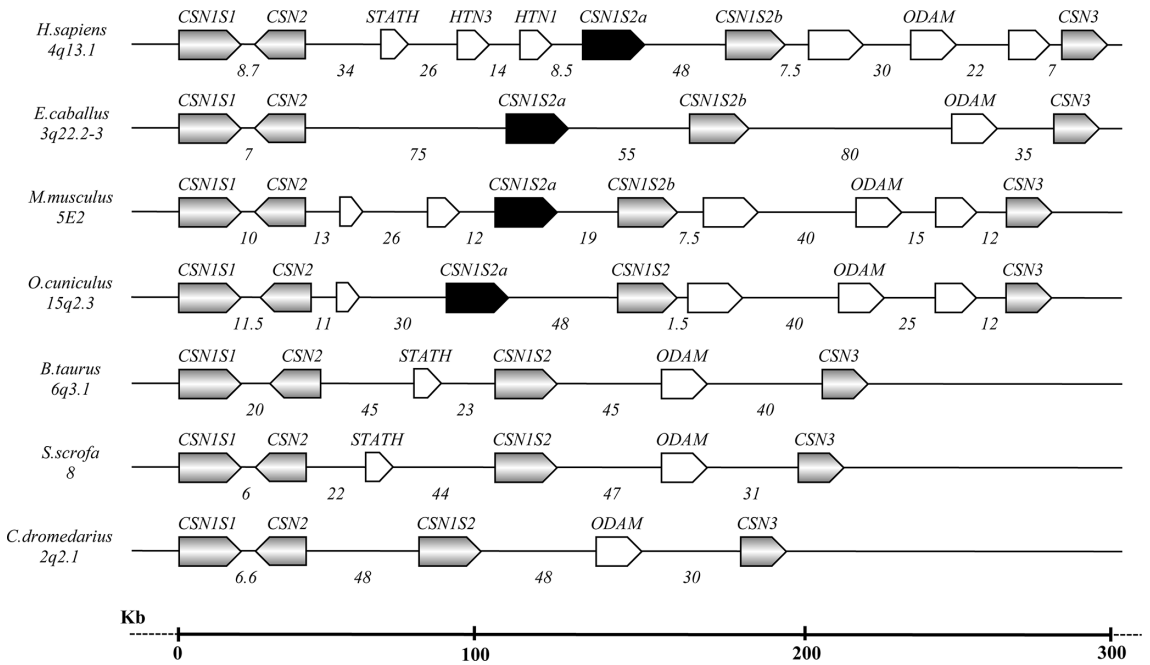


FIGURE 1 | Schematic representation of the casein gene cluster in different species, chromosome location, gene orientation, and related intergenic distances based on comparative genomic analysis (Modified from Martin et al., 2013 and Ryskaliyeva et al., 2018).

TABLE 2 | Positions, sizes, and exon numbers of the casein genes and related intergenic distances occurring in the cluster. The contig 039344 available in EMBL with the acc. no. AGVR01039100.1 and isolated from the whole genome sequence of the feral camel (Bactrian Camels Genome Sequencing and Analysis Consortium) was used as reference. D, Dromedary; C, other camelids; L, Llama (Pauciullo and Erhardt, 2015).

Gene	Position	Size (bp)[A]	Intergenic distance (bp)[B]	Total size (bp)[A+B]	Exons
CSN1S1	242,112 to 258,587	16,476			20[D]/21[C]
↓			6,600		
CSN2	265,187 to 273,094	7,908			9
↓			48,261		
CSN1S2	321,355 to 335,898	14,544			17
↓			85,699		
CSN3	421,597 to 430,955	9,359			5[D,C]/6[L]
				188,847	

of the complete DNA interval (*CSN1S1* to *CSN3*) is 34%, and the average repeat content is about 20% (**Supplementary Table 1**).

Single Nucleotide Polymorphisms and Microsatellites

The comparative analysis of the complete casein cluster (~190kb) showed a total of 12,818 SNPs (**Table 3**). For all camelids, the most polymorphic gene was the *CSN1S1* (99 SNPs in Bactrian *vs*. 248 in dromedary *vs*. 626 in alpaca), whereas the less polymorphic gene was the *CSN3* in Bactrian (22 SNPs) and alpaca (301 SNPs) and the *CSN1S2* in dromedary (79 SNPs).

The analysis of sequences for microsatellites discovery found a total of 35 microsatellites. Fifteen were identified in all camelids, six were shared among three species, seven were in common between two species, whereas seven were found to be species-specific (**Table 4**). No specific microsatellites were found in the wild feral camel.

Interspersed Elements

The analysis of the interspersed repeats evidenced a total of 696 elements (169 in wild camel, 171 in Bactrian, 174 in dromedary, and 182 in alpaca). Most of them (82.3%; 572 elements) were common among all camelids; 84 interspersed repeats (12.0%) were shared between two (18 elements) or three species (66 elements), whereas 40 interspersed elements (5.7%) were species-specific (**Supplementary Table 1**). The diagrammatic representation of interspersed elements is reported in **Figure 2**.

Alpaca showed 27 species specific transpositions, whereas the Old word camelids were poor in these elements, only 12 in total. In particular, four interspersed elements, mainly belonging to LINE-1 (L1) retrotransposons, were found in the wild feral; three repetitive sequences (MER35, ERV_36_MD_I, RLTR25_MM) were detected in the Bactrian; and five specific repetitive

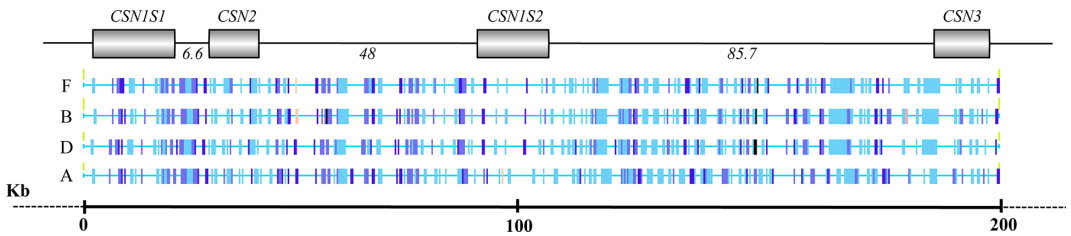


FIGURE 2 | Overview of the interspersed elements identified in the casein gene cluster in wild feral camel **(F)**, Bactrian **(B)**, dromedary **(D)**, and alpaca **(A)**.

elements (CHARLIE8; miR; LTR2_Vpa; SloEFV-I; L1MAB2_ML) were found in dromedary (**Table 5**).

Promoters

The analysis of the casein gene promoters for the discovery of the consensus sequences for transcription factors evidenced 505 binding sites with a score ranging between 85% and 100%. The most representative elements correlated to protein and milk production were those belonging to the Oct family (octamer-binding elements), GATA-binding proteins, CCAAT-enhancer-binding proteins (C/EBPs), and broad activators like AP-1, AP-2, SP1, etc. The consensus sequences common to the four caseins and showing the highest binding scores are reported in **Table 6**.

Experimental Data

Genotype and Haplotype Analyses in Dromedaries

Two dromedary populations (Sudanese and Nigerian) were genotyped for SNPs located on three genes (c.150G > T, *CSN1S1* allele C; g.2126A > G, *CSN2* promoter; and g.1029T > C, *CSN3* promoter), known for being polymorphic. **Figure 3** shows the genotype pattern for the three polymorphisms.

The allelic frequencies are reported in the **Table 7**. The allele C at *CSN1S1* is very rare in Nigerian (0.054) and Sudanese dromedaries (0.094), whereas the frequency of the allele G at *CSN2* is almost inverted (0.550 in Nigerian *vs*. 0.350 in Sudanese), as happens also for the allele C at *CSN3* (0.549 in Nigerian *vs*. 0.377 in Sudanese). No deviation from Hardy-Weinberg equilibrium was found for all loci within populations.

On the basis of the genotypes detected for each *locus*, eight haplotypes were observed in both populations (**Table 7**). Sudanese camels showed a higher frequency (0.348) of the haplotype GAC compared to the Nigerian (0.187), where the most represented haplotype (0.290) was GGC, rather underrepresented in the Sudanese camels (0.028). Overall, the haplotype GAC was the most frequent (0.288), whereas TGC was the rarest (0.005).

Cytogenetic Mapping

The investigated alpacas were karyotyped, and the analysis of the RBA-banding pattern showed karyologically normal animals (2n = 74, XX).

Five PCR amplicons spanning the casein *loci* were mixed together and used to set up a fluorescence *in situ* hybridization (FISH) based method for the mapping on alpaca chromosomes (**Figure 4**). The specificity of the amplified probes was first assessed by agarose gel electrophoresis and then by Sanger sequencing. The comparison with the feral camel genome sequence (EMBL acc. no. AGVR01039100.1) and with the homologous camel *CSN2* (EMBL acc. no. HG969421) and *CSN3* (EMBL acc. no. HE863813) gene sequences confirmed that the probes belonged to the casein genes.

The reliability of the gene signal detection by FISH was assessed on 30 counted metaphases. The FISH efficiency was 92.2% on average

TABLE 3 | Number of polymorphic sites differentiated in substitutions (Sub), insertions (Ins), and deletions (Del) found in the casein cluster region of Bactrian, dromedary, and alpaca by the comparative genomic analysis using the wild camel sequence (AGVR01039100.1) as reference, including the total numbers of polymorphic sites (TOT) within gene by species and in total.

	C. bactrianus				C. dromedarius				V. pacos				TOT
	Sub	Ins	Del	TOT	Sub	Ins	Del	TOT	Sub	Ins	Del	TOT	
CSN1S1	17	16	66	**99**	50	141	57	**248**	333	248	45	**626**	**973**
Intergenic 1	6	0	0	**6**	23	3	0	**26**	206	15	23	**244**	**276**
CSN2	7	0	90	**97**	36	87	91	**214**	216	8	141	**365**	**676**
Intergenic 2	70	32	379	**481**	216	156	378	**750**	1085	544	825	**2,454**	**3,685**
CSN1S2	24	15	17	**56**	41	28	10	**79**	360	74	75	**509**	**644**
Intergenic 3	56	27	105	**188**	234	84	99	**417**	1406	419	342	**2,167**	**2,772**
ODAM	5	1	2	**8**	30	13	4	**47**	172	17	36	**225**	**280**
Intergenic 4	92	165	51	**308**	148	138	142	**428**	722	80	1405	**2,207**	**2,943**
CSN3	15	4	3	**22**	38	65	143	**246**	220	16	65	**301**	**569**
TOT	**292**	**260**	**713**	**1,265**	**816**	**715**	**924**	**2,455**	**4,720**	**1,421**	**2,957**	**9,098**	**12,818**

TABLE 4 | List of microsatellites found in the casein cluster of camelids. Short tandem repeats polymorphic among the species are in italics. Species-specific microsatellites correspond to gray cells. Positions are indicated according to the corresponding GenBank sequence (wild feral: AGVR01039100.1; Bactrian: NW_011517196.1; dromedary: NW_011591251; alpaca: KN269544); therefore, they are complementary (Compl) for Bactrian and dromedary.

Wild feral				Bactrian				Dromedary				Alpaca			
Position	**Cicle**	**Repeats**	**Unit**	**Position Compl**	**Cicle**	**Repeats**	**Unit**	**Position Compl**	**Cicle**	**Repeats**	**Unit**	**Position**	**Cicle**	**Repeats**	**Unit**
254698	2	7	TC	7090479	2	7	TC	491968	2	7	TC	353736	2	7	TC
254760	*2*	*19*	*TG*	*7090418*	*2*	*27*	*TG*	*491907*	*2*	*26*	*TG*	*353797*	*2*	*17*	*TG*
257864	2	6	TG									356896	2	6	TG
274229	*2*	*8*	*TA*	*7071025*	*2*	*10*	*TA*	*472362*	*2*	*8*	*TA*	*373152*	*2*	*7*	*TA*
				7069900	3	10	TA								
275356	*3*	*7*	*TAT*					*471241*	3	6	TAT				
293407	2	6	TA	7051865	2	6	TA	453324	2	6	TA	391869	2	6	TA
294330	*3*	*9*	*TAT*									*392786*	*3*	*15*	*TAT*
294361	*3*	*9*	*ATC*	*7050923*	*3*	*10*	*ATC*	*452355*	*3*	*9*	*ATC*	*392832*	*3*	*10*	*ATC*
								441188	4	6	ATTG				
327761	*4*	*7*	*AGAC*	*7017880*	*4*	*7*	*AGAC*	*419012*	*4*	*10*	*AGAC*				
												446456	3	10	TCC
348071	3	7	CCG	6997586	3	7	CCG					446668	3	7	CCG
348583	2	6	TC	6997084	2	6	TC	398200	2	6	TC				
349673	*2*	*9*	*TG*	*6995994*	*2*	*11*	*TG*	*397110*	*2*	*7*	*TG*	*448262*	*2*	*6*	*TG*
349701	*2*	*6*	*TA*	*6995962*	*2*	*6*	*TA*	*397086*	*2*	*6*	*TA*	*448282*	*2*	*10*	*TA*
362062	*4*	*10*	*TAGA*	*6983602*	*4*	*9*	*TAGA*	*384718*	*4*	*12*	*TAGA*				
362796	2	6	AT	6982856	2	6	AT	383956	2	6	AT	461613	2	6	AT
												461638	2	6	TA
369500	2	7	CA	6976226	2	7	CA	377270	2	7	CA	468360	2	7	CA
373342	2	9	TC	6972383	2	9	TC								
373360	*2*	*12*	*AC*	*6972365*	*2*	*16*	*AC*	*373405*	*2*	*14*	*CA*	*472211*	*2*	*7*	*CA*
												488423	2	6	AC
												495913	2	6	TA
397130	2	6	AT	6948637	2	6	AT	349656	2	6	AT	495926	2	7	AT
				6943100	4	10	TAAC	344115	4	9	TAAC				
404226	*2*	*7*	*CA*	*6941532*	*2*	*7*	*CA*	*342508*	*2*	*11*	*CA*	*502998*	*2*	*15*	*CA*
406619	*2*	*12*	*TG*	*6939139*	*2*	*12*	*TG*	*340113*	*2*	*14*	*TG*	*505398*	*2*	*8*	*TG*
408958	*3*	*8*	*ATG*	*6936762*	*3*	*11*	*ATG*	*337850*	*3*	*8*	*ATG*	*507654*	*3*	*6*	*ATG*
412396	*2*	*8*	*AC*	*6933319*	*2*	*13*	*AC*	*334433*	*2*	*11*	*AC*	*511052*	*2*	*16*	*AC*
415959	2	6	TA	6929747	2	6	TA	330866	2	6	TA				
								330472	2	8	TG				
416355	*2*	*10*	*TG*	*6929351*	*2*	*10*	*TG*	*330404*	*2*	*7*	*TG*				
416377	2	6	CA	6929329	2	6	CA								
				6914165	*2*	*11*	*AC*	*315383*	*2*	*10*	*AC*				

TABLE 5 | Species-specific interspersed elements found by the comparative genomic analysis of the casein cluster in camelids (A, alpaca; B, Bactrian; D, dromedary; F, feral camel) and listed in order 5' > 3' as they appear in the cluster. Repbase was used as repeat database. Positions are indicated according to the corresponding GenBank sequence (wild feral: AGVR01039100.1; Bactrian: NW_011517196.1; dromedary: NW_011591251; alpaca: KN269544). The direction of the repeat fragment is indicated as d = direct or c = complementary; Sim, similarity with repeat element in the database; and Pos/Mm, Ts column is a ratio of mismatches to transitions in nucleotide alignments. The closer Pos/Mm, Ts number is to 1 the more likely is that mutations are evolutionary.

Species		Position	Name	Class	Dir	Sim	Pos/Mm : Ts	Size (bp)
A		345745-345817	ERV3-1_SSc-I	ERV	c	0.785	2.00	72
		352418-352502	ERV44_MD_I	ERV	c	0.712	1.47	84
D	Compl. to	485496-485356	CHARLIE8	DNA/hAT	c	0.647	1.59	140
	Compl. to	478096-478030	miR	SINE	c	0.806	1.22	66
F		268407-268493	THER2	SINE	c	0.755	2.00	86
A		369205-369268	MER4BI	ERV	c	0.822	2.00	63
		372430-372493	L1-2B_EC	LINE	d	0.790	1.50	63
		381127-381183	MARINER4_MD	Mariner	d	0.836	1.00	56
B	Compl. to	7053559-7053474	MER35	MER	d	0.717	1.58	85
A		391846-391888	L1A-2_MD	LINE	d	0.809	1.17	42
		395574-395643	Zaphod3	DNA/hAT	d	0.760	1.18	69
		408816-408940	ERV3-5-EC_LTR	ERV	c	0.830	2.11	124
F		310230-310295	L1-2_Vpa	LINE	d	0.750	1.67	65
A		413499-413572	HAL1-3_ML	LINE	c	0.783	1.50	73
		417674-417832	L2	CR1	d	0.652	1.47	158
		417888-418114	L2	CR1	d	0.654	1.83	226
D	Compl. to	426349-425998	LTR2_Vpa	ERV	d	0.758	4.61	351
A		422684-422764	LTR28_OC	ERV	c	0.707	1.62	80
		426397-426425	SQR2_MM	Sat	d	0.931	2.00	28
		427755-427846	L1-2_Vpa	LINE	d	0.768	1.80	91
		451056-451088	HERVK3I	ERV	d	0.882	1.50	32
D	Compl. to	383725-383659	SloEFV-I	ERV	d	0.776	1.83	66
A		465703-465770	MER104B	DNA	c	0.753	1.78	67
		468337-468372	ERV2-1-I_Opr	ERV	c	0.865	1.00	35
		471920-471998	RMER3D-int	ERV	d	0.782	1.83	78
		478503-478578	LTR16	ERV	c	0.701	1.69	75
		480865-480899	SINE_VV	SINE	c	0.888	1.50	34
F		393221-393292	L1-1H_Cpo	LINE	d	0.833	2.00	71
A		493574-493639	MER28	Mariner	d	0.791	1.44	65
		502950-503003	RLTR17B_Mm	ERV	c	0.763	1.33	53
		505352-505434	RLTR17_MM	ERV	d	0.765	1.56	82
B	Compl. to	6936860-6936781	ERV36_MD_I	ERV	d	0.779	1.71	79
A		509996-510114	PRIMA4_I	ERV	d	0.736	1.85	118
B	Compl. to	6933341-6933305	RLTR25_MM	LTR	c	0.8611	2.00	36
F		412370-412416	L1-3_TS	LINE	c	0.847	9.90	46
A		519191-519256	L1MdV_II	LINE	c	0.761	1.44	65
D	Compl. to	328118-328009	L1MAB2_ML	LINE	d	0.684	1.48	109
A		519318-519379	ERV2-3_STr-I	ERV	d	0.796	1.50	61
		524222-524291	UCON28c	Int. Rep.	c	0.843	2.50	69

TABLE 6 | Most representative consensus motifs for transcription factors detected in the 5'-flanking regions of camelids by TFBIND software and present in all caseins with higher binding score (BS). DNA strands (S) in direction 5' > 3' are indicated by +. The opposite strands are indicated by -.

Transcription factor	Consensus motif	*CSN1S1*			*CSN2*			*CSN1S2*			*CSN3*		
		Position	S	BS	Position	S	BS	Position	S	BS	Position	S	BS
AML1/Runx	TGTGGT	-259/-254	-	0.873	-57/-52	-	1.000	-300/-295	+	0.910	-61/-56	-	0.850
AP-1	RSTGACTNMNW	-186/-176	-	0.850				-98/-88	+	0.851	-104/-94	-	0.890
C/EBP	NNTKTGGWNANNN	-304/-292	-	0.940	-271/-259	+	0.911	-51/-39	-	0.927	-58/-45	-	0.875
GATA	NNNGATRNNN	-106/-97	+	0.870	-183/-174	+	0.887	-350/-341	-	0.860	-124/-115	-	0.931
HNF3	NNNTRTTTRYTY	-83/-72	+	0.880	-77/-66	+	0.928	-338/-327	+	0.928	-20/-9	+	0.932
MyoD	SRACAGGTGKYG				-307/-296	+	0.874	-265/-254	+	0.925	-62/-51	-	0.858
Oct-1	NNNRTAATNANNN	-267/-255	-	0.929	-131/-120	+	0.917	-186/-174	+	0.947	-84/-72	+	0.852
Pbx-1	ANCAATCAW	-45/-37	+	0.942	-109/-101	+	0.903	-221/-213	+	0.899	-33/-25	-	0.912
SRY	AAACWAM	-253/-248	+	0.941	-192/-186	-	0.960	-169/-163	+	0.947	-14/-8	-	0.939
MGF/STAT5	TTCCCRKAA	-270/-262	+	0.931	-94/-86	-	0.870	-292/-284	-	0.956			
TATA-box	WTATAAAW	-31/-25	+	0.980	-28/-19	+	0.910	-23/-16	+	0.865	-18/-11	-	0.927

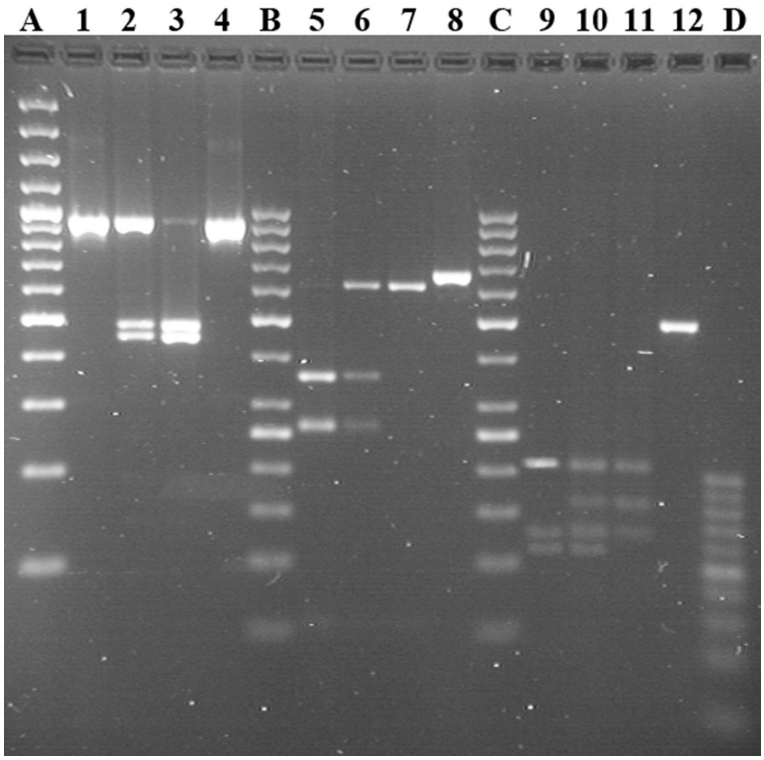


FIGURE 3 | Genotyping of *CSN1S1*, *CSN2*, and *CSN3* by PCR-RFLP in Sudanese and Nigerian dromedary populations. Lines 1–4: locus *CSN1S1* c.150G > T; genotypes TT, GT, and GG reported in lines 1, 2, and 3, respectively. Lines 5–8: locus *CSN2* g.2126A > G; genotypes GG, AG, and AA reported in lines 5, 6, and 7, respectively. Lines 9–12: locus *CSN3* g.1029T > C; genotypes CC, TC, and TT reported in lines 9, 10, and 11, respectively. Lines 4, 8, and 12 show undigested PCR products each belonging to the relative locus. Line A shows the GeneRuler™ 100 bp plus DNA Ladder (Thermo Scientific). Lines B and C show GeneRuler™ 50 bp DNA Ladder (Thermo Scientific). Line D shows 20bp DNA Ladder (Jena Bioscience).

TABLE 7 | Allele and haplotype frequencies detected for the SNPs c.150G > T, g.2126A > G, and g.1029T > C at the casein loci in Sudanese and Nigerian dromedary populations.

	Allele frequencies							
	CSN1S1 **c.150G > T**		***CSN2*** **g.2126A > G**				***CSN3*** **g.1029T > C**	
	G	**T**	**A**	**G**			**T**	**C**
Sudanese (n = 198)	**0.906**	**0.094**	**0.650**	**0.350**			**0.623**	**0.377**
Shanbali	0.900	0.100	0.640	0.360			0.540	0.460
Khali	0.921	0.079	0.723	0.277			0.700	0.300
Arabi	0.942	0.058	0.704	0.296			0.654	0.346
Lahaoi	0.888	0.112	0.587	0.413			0.607	0.393
Nigerian (n = 69)	**0.946**	**0.054**	**0.450**	**0.550**			**0.451**	**0.549**
	Haplotype frequencies							
	1	2	3	4	5	6	7	8
	GAC	**GAT**	**GGC**	**GGT**	**TAC**	**TAT**	**TGC**	**TGT**
Sudanese	**0.348**	**0.263**	**0.028**	**0.269**	**0.007**	**0.026**	**0.004**	**0.052**
Shanbali	0.415	0.126	0.003	0.354	0.019	0.067	0.000	0.013
Khali	0.277	0.401	0.039	0.203	0.016	0.037	0.001	0.023
Arabi	0.254	0.374	0.071	0.241	0.024	0.019	0.004	0.009
Lahaoi	0.347	0.239	0.039	0.269	0.001	0.001	0.015	0.087
Nigerian	**0.187**	**0.226**	**0.290**	**0.223**	**0.014**	**0.029**	**0.006**	**0.023**
Over-all frequency	**0.288**	**0.271**	**0.099**	**0.253**	**0.011**	**0.027**	**0.005**	**0.042**
Standard Error	0.018	0.021	0.015	0.019	0.006	0.007	0.004	0.008

In bold the the allele and haplotype frequencies for the complete Sudanese and Nigerian populations analysed, as well as for the over-all frequency.

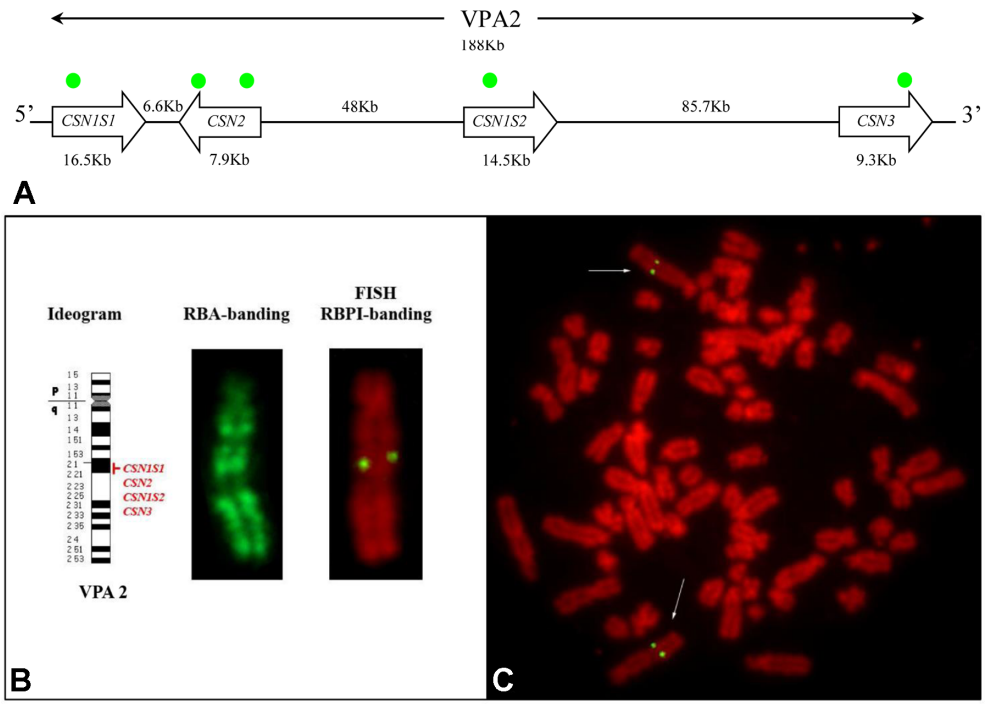


FIGURE 4 | Casein cluster organization and genes mapping in alpaca (VPA). **(A)** Schematic representation of the genomic arrangement of the casein loci along the chromosome. The approximate position of the gene-specific probes used for the mapping is indicated with green dots. **(B)** Diagrammatic representation of VPA 2, RBA-banding, and RBPI staining. (**C**) Fluorescence *in situ* hybridization obtained by using *CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3* gene probes shows specific signal on chromosome 2q2.1. Fluorescein isothiocyanate signals were superimposed on RBPI-banding (R-banding using early BrdU-incorporation and propidium iodide staining).

(range 86–97%). Two couple of symmetrical spots, each belonging to the sister chromatids of the two homologous chromosomes, were identified in the analyzed R-banding metaphases (**Figure 4B**, **C**). The application of the propidium iodide-staining (RBPI-FISH) allowed the mapping of the casein genes to the chromosome 2q21. Alpaca chromosome 2 is reported in detail in **Figure 4B**.

No further hybridization signals were detected on the other chromosomes, thus confirming the cluster organization of casein genes with no duplications (**Figure 4A**).

DISCUSSION

The dramatic progress of sequencing technologies and the enormous reduction in the cost of sequencing opened the era of the genomics. Genome sequencing projects provided a huge amount of data, but, despite the new research abilities have been developed, new problems are also coming out. For example, the low coverage assembly and the tentative annotations often built on the human genome, led to repetitive information, exon losses, and errors in gene annotations. This is still more evident for less explored species like those belonging to *Camelidae*. For instance, the *CSN2* in the alpaca genome has been annotated without the exon 3, because this exon is out-spliced in human that was used as comparative reference genome. However, the DNA sequence of the exon 3 can be found about 130 bp upstream of the provided *Vicugna pacos* genomic sequence (Pauciullo and Erhardt, 2015). Furthermore, although the genome sequencing has been completed for the wild feral, dromedary, and Bactrian camels, as well as for the alpaca (Wang et al., 2012; Wu et al., 2014; Fitak et al., 2016), their annotation is still incomplete. Therefore, it is necessary to gain more data to help the annotation in camelids.

In this context, we focused our investigation on genes encoding the main component of milk proteins, providing for the first time a detailed comparative analysis of the casein cluster in camelids; information on haplotype variability in two dromedary populations; and the physical map of the casein genes in alpaca.

Multiple Bioinformatics Analysis

Cluster Organization

Milk proteins and the corresponding coding genes have been deeply studied because they represent in all species the primary source of nutrients for the new born. Caseins (αs1, β, αs2, and κ) are the main component of milk proteins, and they are coded by single autosomal genes (*CSN1S1, CSN2, CSN1S2,* and *CSN3,* respectively) clustered in a DNA stretch closely linked.

In camelids, the entire casein cluster covers a region slightly less than 190 kb (**Table 2**), and it appears to be "contracted" compared with the cluster observed in other species, in general spread between 250 and 350 kb, depending on the species (Martin et al., 2013; Ryskaliyeva et al., 2018). For instance,

the human casein gene cluster is characterized by 11 genes, which exist approximately in the same position in other species. Conversely, the same DNA region in camelids only includes five genes (**Figure 1**).

It is known that the genome expansion is a key mechanism of diversification in the evolution, with gene duplication, exon duplication, and alternative splicing acting as the major driving forces. Jones et al. (1985) proposed that the casein genes evolved from an ancestral gene by a combination of intra- and inter-genic exon duplications. In this respect, some mammals, including horse, donkey, rodents, and rabbit, show two αs2-casein encoding genes (*CSN1S2*a and *CSN1S2*b), which may have arisen by a relatively recent gene-duplication event and may represent examples of paralog duplication (Stewart et al., 1987; Ginger et al., 1999; Cosenza et al., 2010). Conversely, the analysis of the sequences in the intergenic region *CSN2* - *CSN3* of camelids did not evidence the existence of a second *CSN1S2* gene (**Figure 1**). This finding confirms the phylogenetic data of Rijnkels (2002), which demonstrated the gene loss in the *Artiodactyla*, while further divergence between the two gene copies in the other species was partially achieved by differential exon usage. Conversely, the *ODAM* gene was found in camelids and, similarly to the casein genes, it is always present in all species (**Figure 1**).

The organization and the orientation of the casein genes in camelids were conserved as in the other species studied to date (Fujiwara et al., 1997; Rijnkels et al., 1997; Milenkovic et al., 2002; Pauloin et al., 2002; Rijnkels, 2002; Ramunno et al., 2004) (**Figure 1**). Also in camelids, the conservation in the orthologues casein genes is mainly in the 5' UTR and the signal peptide. In fact, in all casein genes, the first exon encodes the 5'UTR. In the Ca-sensitive caseins (*CSN1S1, CSN2, CSN1S2*), the second exon carries the remaining 12 nucleotides of the 5'UTR and encodes the signal peptide and two amino acids of the mature protein. In the *CSN3* gene, the signal sequence is encoded by the exon 2 and part of exon 3.

The comparative analysis of the genome sequences with the casein transcripts in llamas (Pauciullo and Erhardt, 2015) and dromedary (Kappeler et al., 1998) and the sequences of dromedary *CSN2* (EMBL acc. No. HG969421) and *CSN3* (EMBL acc. no. HE863813) genes allowed the identification of the exons. The architecture of the four genes is extremely "fragmented" in terms of coding regions. The dromedary *CSN1S1* consists of 20 exons, whereas the other camelids shared an organization in 21 exons. The main difference is due to the exon 20, taking as reference the llama cDNA reported by Pauciullo and Erhardt (2015). This exon (44 bp long), partially coding for the termination stop codon (exon 19 5'-TG … A-3' exon 20), was not found by Kappeler et al. (1998). The reason lies in the mutation that occurred at the donor splice site of dromedary sequence (**Supplementary Figure 1**), which alters the correct identification of the splicing sites and skips out the exon. The correct reading frame is then restored by the next exon, which starts also with an adenine, thus restoring the termination stop codon. Conversely, in the other camelids, the exon 20 and the corresponding splicing sites are conserved. In addition, the analysis of its sequence showed an identity of 95% with the exon 18 of the *CSN1S1* gene (EMBL acc. no. EU025875.1) and cDNA in pig (NM_001004029), and a similarity of 91% with bovine (X59856), goat (AJ504710), and sheep (JN560175) homologous exon.

The *CSN2* gene is conserved in the structure (nine exons) and in the inverted orientation, also in comparison with other species (**Figure 1**).

The *CSN1S2* is arranged in 17 exons in all camelids. This structure confirms the data already published for llamas (Pauciullo and Erhardt, 2015), but it is slightly different from the information reported previously for dromedary (Kappeler et al., 1998). These authors based their study on a reverse approach, from protein to mRNA. The clone library was screened by degenerated primers, whose sequences were deduced from the sequencing of tryptic peptide digestions. Furthermore, they never mentioned the number of clones analyzed; therefore, it is likely that not all mRNA populations were found. To date, no other studies on casein transcripts were carried out in dromedary, but the comparison of the *CSN1S2* genome sequences with the corresponding llama cDNA (Pauciullo and Erhardt, 2015) evidenced that the exon 12 is conserved also in the other camelids, and no mutations affect splicing elements. This exon that is 27 bp long and coding for a peptide of nine amino acids (ENSKKTVDT) is homologous (96.3%) to the predicted αs2 cDNA of *Pantholops hodgsonii* (XM_005985429), a wild Tibetan antelope well adapted to survive in severe conditions, which remind analogous situation of camelids. The same exon shared an identity of 90% with the exon 13 of the bovine (M94327.1) and goat (AJ297316.1) *CSN1S2* gene; about 89% with buffalo (FM865619), sheep (GU169085), and horse (NM_001170767) homologous cDNA; and 85% with the exon 14 of donkey (FN298386.2) *CSN1S2* I gene. Therefore, we postulate that the exon 12 was not described in dromedary *CSN1S2* gene because it was likely spliced out in the pool of clones as analyzed by Kappeler et al. (1998). Recently, Ryskaliyeva et al. (2018) reported a deep characterization of milk protein in Old World camelids accomplished by LC-ESI-MS. However, these authors reported a different phosphorylation level of αs2-CN and did not analyze deeply the primary transcripts. More recently, an extensive protein characterization proposed by the same authors confirmed the splicing of this nine amino acids in the Old World camelids and evidenced new αs2-CN isoforms (Ryskaliyeva et al., 2019). Considering these recent findings, a deep investigation at transcript level is highly beneficial to elucidate all constitutive and alternative splicing events in mRNA maturation process of *CSN1S2*.

Also, the structure of the *CSN3* is different between the Old and New World camelids. The gene arrangement in five exons is very well conserved among the species (Rijnkels, 2002); however, in llamas, 66.6% of the *CSN3* gene transcripts showed one additional "cryptic" exon of 43 bp (Pauciullo and Erhardt, 2015). This extra exon was not identified in the dromedary *CSN3* transcripts (Kappeler et al., 1998), although the sequence is present in the corresponding intron of all camelids *CSN3* gene sequences. Although many nucleotide differences discriminate the intron of both Old and New World camelids, the computational analysis of spliceosome specific sites confirmed the occurrence of the splicing elements: a branch point, a polypyrimidine tract, and a terminal AG acceptor site (score 0.87) at the extreme 3' end of the intron 1. Moreover, the occurrence of a donor site at the 5'

end of the intron 2 was estimated with a score of 0.99 confirming the existence of the additional exon (**Supplementary Figure 2**).

The occurrence of cryptic exons in the casein genes was already observed in camelids. For instance, in the *CSN1S1* gene of dromedary and llamas, the out-splicing of the exon 18 generates two variants (A and B) differing for the octapeptide (EQAYFHLE). Additional examples exist also in other species. For example, the exon 3 of the human β-casein was described as cryptic due to the interruption of the polypyrimidic tract of the intron 2 by four purines (Menon et al., 1992). In camelids, the occurrence of a larger polypyrimidine DNA tract and the existence of different branch points (BPs) besides the conventional mammalian BP sequence (5'-YTRAY-3') might be the reason of an alternative skipping of the cryptic exon, as already observed in llamas (Pauciullo and Erhardt, 2015).

The presence of an extra exon in *CSN3* cDNA would also add a new reading frame, *de facto* extending the length of the signal peptide of six amino acids (*MLLGAI*) at NH_2-terminus (three coming from the cryptic exon 2 and three coming from the reading frame of the following exon). In addition, two possible translation start codon (ATG) would occur (one on the cryptic exon 2 and one canonical on the exon 3), without altering the normal reading frame. Therefore, both protein variants would have a "functional" signal peptide to guarantee the crucial role of κ-CN in the casein micelle maintenance (**Supplementary Figure 2**).

Despite the new information provided by our analysis, a deep investigation at transcript level is needed, at least in dairy camels (dromedary and Bactrian), to elucidate the existence of a real difference in the *CSN3* gene structure between Old (five exons) and New World (six exons) camelids and to clarify the other issues stressed in the present study.

Single Nucleotide Polymorphisms

The comparative analysis of the casein cluster among the four investigated camelids showed a high level of genetic diversity. Considering only simple events (nucleotide substitution, insertions, and deletions), a total of 12,818 SNPs were found (**Table 3**).

It is known that genetic polymorphisms contribute to variations in phenotypes. In ruminants, the *CSN1S1* can be surely considered as the most polymorphic gene among caseins (Ramunno et al., 2005; Caroli et al., 2006; Caroli et al., 2009). Although few studies were carried out in camelids, our data confirm the highest level of variation in *CSN1S1* gene (99 SNPs in Bactrian *vs.* 248 in dromedary *vs.* 626 in alpaca), with a total of 24 SNPs occurring in the exons, 13 of which falling in translated regions and, therefore, responsible of amino acids variations among the species.

Currently, at least four protein variants (A, B, C, and D) were detected in the dromedary αs1-CN, and the molecular event responsible for these phenotype variations was clarified in three out of four cases. The alternative out-splicing of the exon 18 differentiates the variant A (207 aa) and B (215 aa). This genetic event is due to the insertion of 11 bp (ATTGAATAAAA) in the intron 17, which negatively affects the secondary structure of the *CSN1S1* A pre-mRNA for the creation of a hairpin and coiled loop (Erhardt et al., 2016). Conversely, the allele C is due to a single nucleotide polymorphism (c.150G > T; GenBank ID JF429138) occurring at the exon 5 and resulting in the amino acid substitution p.30Glu > Asp in the mature protein (Shuiep et al., 2013). Recently, Erhardt et al. (2016) reported a new variant, named D. This showed a different IEF profile, but the analysis of the gene sequence did not evidence any substantial difference with the A allele, apart from an insertion of 11bp in the intron 17. Therefore, the molecular event is still unknown. The *CSN1S1* gene is also polymorphic in llamas, where four protein variants corresponding to four haplotypes were recently reported by Pauciullo et al. (2017). Also in this case, the molecular bases of the differences were identified in two SNPs (c.366G > A, exon 12 and c.690C > T, exon 19) responsible for the amino acid substitutions p.86Val > Ile and p.194His > Tyr, respectively.

The comparison of *CSN2* genes showed 676 polymorphic sites. This gene was the second most polymorphic in camelids with 35 mutations realized in the coding regions, of which 16 occurring in translated regions. The *CSN2* was well characterized in ruminants (Caroli et al., 2006; Cosenza et al., 2007; Caroli et al., 2009), and a detailed description of the gene was also reported in dromedary and Bactrian camels (Pauciullo et al., 2014). To date, only one polymorphism affecting the protein was reported in Bactrians, that is the SNP c.666G > A, which is responsible for the amino acid change p.201Met > Ile (Pauciullo et al., 2014), recently confirmed by Ryskaliyeva et al. (2018). SNP discovery in dromedary highlighted two single polymorphisms: the SNP g.4175C > A that occurred within the codon 7 of the signal peptide, but it was a synonymous mutation (GCC^{Ala} > GCA^{Ala}); and the SNP g.2126A > G that occurred in TATA-box of dromedary *CSN2* promoter and was more interesting because it putatively affects the transcription factor binding activity (Pauciullo et al., 2014).

A similar number of genetic markers (644) were also found by the comparative analysis of the *CSN1S2* genes. This gene resulted the least polymorphic in dromedary with 79 SNPs in total (**Table 3**). The SNPs affecting the exons were 27, of which 17 occurred in translated regions and, therefore, putatively responsible of amino acids differences. As for the other casein fraction, also the *CSN1S2* was well studied in ruminants, and many alleles were found (Caroli et al., 2006; Cosenza et al., 2007; Caroli et al., 2009). Exons skipping are also considered as frequent events for the αs2-casein encoding gene in different species (Boisnard et al., 1991; Bouniol et al., 1993; Cosenza et al., 2009). However, to date, no genetic variants or alternative transcripts have been reported in camelids. Considering the origin and the structure of the *CSN1S2* gene (Rijnkels, 2002), at least rearrangements resulting from alternative splicing of mRNA are expected. Therefore, surely in camelids, this casein gene deserves more attention at gene transcript level.

The least polymorphic gene in camelids was the *CSN3*, with 569 markers found by the comparative analysis (**Table 3**). Twenty-one SNPs were found in the exons, and 14 of them occurred in the translated regions, mostly located in the exon 4 (12 SNP found only in Alpaca). The *CSN3* is not evolutionarily related to the Ca-sensitive casein genes, but is physically linked to this gene family, and is functionally important for stabilizing the Ca-sensitive caseins in the micelle. Therefore, mutations occurring in this gene can be particularly important for the biological role carried out by the κ-CN. Pauciullo et al. (2013a)

carried out genetic diversity discovery in Sudanese dromedary. However, no polymorphisms were found in the coding regions, whereas the only interesting SNP was found in the promoter region (g.1029T > C) because affecting the consensus site for the transcription factor HNF-1 just upstreams the exon 1. So far, many genetic variants of κ-casein were identified at protein or DNA level in many species. The absence of polymorphisms in *CSN3* coding regions suggests that the level of genetic variations in camel κ-casein is very low in comparison with other species (Caroli et al., 2006; Carneiro and Ferrand, 2007; Hobor et al., 2008; Caroli et al., 2009). Therefore, a deeper analysis of camel *CSN3* would be necessary to search for genetic diversity, and further studies would be required in order to assess the potential impact on the quali-quantitative properties of camel milk. For example, it is known that dairy cows with the genotype *CSN3* BB produce milk with a significantly higher protein content (Caroli et al., 2009). This led the dairy farmers to select preferentially these cows in order to have a higher cheese yield. Dairy camel breeders could exploit similar advantages, because the presence of quantitative alleles cannot be excluded also in camels. Moreover, Weimann et al. (2009) showed that in cattle *CSN3*, variants are source of different angiotensin I converting enzyme (ACE) inhibitor peptides and revealed their potential role for human health. These bio-functional peptides were found also in camel milk (Al hay and Al Kanhal, 2010). Therefore, it is likely that the genetic variants of the camel κ-casein might also influence its functional role, giving the camel milk an additional value for the human nutrition.

Microsatellites

Despite the progress of genomics and the availability of the high-throughput genotyping platforms, for many domestic species, including camelids, microsatellite analysis still represents a powerful tool for the genetic identification and assessment of parentage analysis in camelids (Penedo et al., 1999; Evdotchenko et al., 2003; Mburu et al., 2003) and characterization of the domestication process of the dromedary (Almathen et al., 2016). The analysis of sequences for microsatellites discovery showed 35 short tandem repeats (**Table 4**). A high level of polymorphism was found among the species (72.7%), demonstrated by 17 microsatellites showing a different number of repeats and seven species-specific short tandem repeats (one in Bactrian, two in dromedary, and four in alpaca), since no over-lapping sequences were found. Although no information is available on allelic diversity, the latter microsatellites can potentially become very useful for species discrimination, genetic diversity, and population structure studies or, simply, for parentage assignment, which is a service in high demand for the camel racing industry (Penedo et al., 1999; Spencer et al., 2010). Furthermore, such a panel of markers, together with other microsatellites distributed along the same chromosome (or SNPs that provide complementary information), gives the opportunity to begin the search for QTLs of economic importance.

Interspersed Elements

Transposable elements have played a fundamental role in species diversification, influencing the evolution of mammalian genomes (Bowen and Jordan, 2002). Camel genome contains about 34% of repetitive DNA (Wang et al., 2012; Fitak et al., 2016), mainly belonging to SINEs and LINEs expanded in the genome by a process known as retroposition. Compared to the whole genome, the DNA fragment containing the casein cluster showed a lower level of repetitive DNA (on average 19.8%). However, the transposition process probably happened in a widespread coverage, within and outside the casein genes, as demonstrated by the short distance occurring between the interspersed elements in each of the investigated species (**Supplementary Table 1**). The comparative analysis showed that 94.4% of the repetitive DNA was shared between two or more species of camelids, whereas 39 interspersed elements (5.6%) were species specific (**Table 5**).

These interspersed elements are useful for a better understanding of the divergent evolution of camelids within the *Tylopoda* family. Recently, the genome analysis of camelids elucidated the divergent time of the ancestors of the New and Old World camelids, indicating that the division between Camelini and Lamini occurred in North America about 16.3 Mya (Wu et al., 2014). Considering this divergence time, it is evident that the interspersed elements common to all camelids were already present in the ancestor genome, whereas the repetitive elements typical of each species were introduced after the separation of Camelini and Lamini tribes. Alpaca showed 27 species specific transpositions, whereas only 12 in the Old word camelids (four in wild feral, three in Bactrian, and five in dromedary). Considering that transposition insertions reflect the level of genome size expansion (Liu et al., 2003), the alpaca genome probably underwent to a more intensive extension due to lineage-specific shifts in transposition activity within the last 17 million years of evolution. This is confirmed by the larger size of alpaca genome (2.05 Gb) compared to that of the Bactrian (2.01 Gb) and dromedary (2.01 Gb) (Wu et al., 2014). Since transpositions are considered powerful mutagens at gene level, their impact on phenotypic change and evolution of camelids might be more significant than considered so far. Examples of phenotypic changes for transposition insertions are present also in the casein genes that, also in this respect, represent a very useful model of study. For instance, the allele E of the *CSN1S1* in goats is characterized by the insertion of a truncated LINE of 457 bp in the last exon, which is responsible of a three-fold reduction of transcriptional rate of the corresponding protein (Pérez et al., 1994). Similarly, in cattle, the *CSN1S1* allele G showed a truncated LINE of 371 bp at the exon 19. The interaction between the LINE sequence and the poly(A) sequence of the mature transcript, reduced the mRNA stability causing a rapid degradation of the transcript and a limited protein synthesis efficiency (Rando et al., 1998).

The presence of repetitive DNA within casein genes in dromedary was already evidenced in *CSN2* (Pauciullo et al., 2014) and *CSN3* genes (Pauciullo et al., 2013b). In these studies, LINEs belonging to L1MA family were found to be species specific in comparison to cattle. None of them affected the exon structure; therefore, no influence on mRNA is expected, as well as on protein production. Furthermore, a lower number of repetitive elements were found in dromedary compared to cattle, thus indicating that *Tylopoda* diverged from Ruminantia before

additional transpositions occurred at different times during the divergence of such suborder (Nijman et al., 2002).

Promoters

Five hundred five motifs for transcription factors enhancing and/or repressing the casein gene expression were found. For brevity, **Table 6** reports only the motifs shared by the four casein promoters and showing higher binding scores. The consensus sequences belonging to the octamer-binding family (Oct), GATA-binding proteins, C/EBPs (CCAAT-enhancer-binding proteins), and ubiquitous activators like Sp1, Ap1, and Ap2, were found more frequently because they are closely linked to protein and milk production.

In particular, 33 C/EBP motifs, 11 mammary gland factor/STAT5 (MGF/STAT5), and 63 octamer-binding protein (Oct-1) were found. These elements initiate the transcription through sinergyc interactions with other motifs (Wyszomierski and Rosen, 2001). For instance, Oct-1 and STAT5 are considered as co-activators, and they can stimulate casein gene expression by hormonal induction (Zhao et al., 2002). In addition, Oct-1 can affect acute myeloid leukemia (AML) factors by reducing its inhibitory role in the DNA binding and creating a complex that stimulates the expression of casein genes (Inman et al., 2005).

The activation of casein expression can be mediated also by hepatocyte nuclear factors-3 (HNF3) by a combined action with nearby C/EBP and glucocorticoid elements (GR) (Schild and Geldermann, 1996; Christoffels et al., 1998). Analogous interactions are supposed for the MyoD transcription factor (Jiang and Zacksenhaus, 2002) and for Pbx1 in a synergic action with glucocorticoid receptors (Subramaniam et al., 2003). Many other motifs were found, including Sp1, NF, YY1, etc., as it was already described in previous studies (Kappeler et al., 2003; Pauciullo et al., 2013a; Pauciullo et al., 2014). However, it is remarkable to point out the existence of one SREBP (sterol regulatory element-binding protein) at position (-61/-51) of the *CSN3* promoter. Although the most known function of this transcription factor is the regulation of genes involved in milk fat pathway (Harvatine and Bauman, 2006), Reed et al. (2008) reported also a down regulation role of SREBP in the expression of caseins.

The description of the most occurring motifs regulating the casein gene expression opens the way to functional studies, which will be necessary to evaluate the influence of these elements on the transcriptional regulation of casein genes in camelids.

Experimental Data

Genotype and Haplotype Analysis

The genotyping of 267 dromedaries for the SNPs at *CSN1S1*, *CSN2,* and *CSN3* showed similarities and differences in the allelic frequencies of the two camel populations. At *CSN1S1*, the variant C (c.150T, p.30Asp) had a very low frequency (< 0.1) in both populations, even lower than that reported by Shuiep et al. (2013) (mean frequency of the allele C = 0.158). Furthermore, this variant does not characterize the other camelids, all carrying the guanine (c.150G, p.30Glu) that can be considered as the ancestral condition within the *Tylopoda* family.

The allele C induces the amino acid replacement p.30Glu > Asp evidenced at protein level by IEF and confirmed at DNA level by the SNP c.150G > T (Shuiep et al., 2013). Taking as reference the variant A of the *CSN1S1*, we carried out bioinformatics analysis to predict the effect of the amino acid change in the secondary structure of the protein and to assess whether it could have an impact on its biological function. The analysis showed an evident change in the secondary structure of α-helix that partially turned to β-sheets (**Supplementary Figure 3**). Furthermore, this structural change in the complex affected a wider region of the protein (amino acids 20–50). However, despite the structural change, PROVEAN analysis showed a score of 0.778, which classifies the mutation as neutral. It is known that any modification of the secondary structure of a protein likely means a change also in the final protein form. If this happens, the functionality of the protein may be affected. This is extremely important in a protein complex such as casein micelle, where the Ca-sensible caseins (αs1-; β and αs2-CN) are closely linked and grouped together in a balanced condition kept by the κ-CN. Examples of strong and defective alleles due to "simple" amino acid changes are known in goats (Cosenza et al., 2008), cattle (Caroli et al., 2009), sheep (Giambra et al., 2010), buffalo (Cosenza et al., 2009), etc. Therefore, further studies are necessary to assess the impact of this variant on the micelle stability, as well as on technological properties and nutrition aspects of the dromedary milk and the related dairy products.

A different situation was observed for the other two SNPs analyzed (g.2126A > G, *CSN2*, and g.1029T > C, *CSN3*), which showed inverted allele frequencies in the investigated populations (**Table 7**). In our knowledge, no genetic programs or selection strategies are applied on camels in both countries (Sudan and Nigeria); therefore, such a difference might be indicative of other effects like genetic drift and/or inbreeding. Nowadays, camel population in Nigeria numbers about 300,000 heads (in 1961, they were only 14,000) and no intensive importing flow of live camels (only 1,300 heads) is documented in the years 1961–2016 (www.faostat.org). Therefore, the current allele distribution could be generated by a founder effect during their domestication time, and the lack of gene flow might have played a role in the differentiation of the Nigerian from the much widespread Sudanese population. This assumption should be confirmed by genetic comparisons with other dromedary populations. However, Nigerian dromedaries were investigated using microsatellites and mitochondrial DNA analysis, and genetic diversity has been found in comparison with Australian, Kenyan, and Canarian Islands populations, assuming inbreeding and/or founder effects as possible reasons (Abdussamad et al., 2015).

On the basis of genotypes detected at each *locus*, eight haplotypes were observed in both populations and, overall, three of them (GAC, GAT, GGT) accounted for more than 80% of the observed variability, with the haplotype GAC most represented (0.288). The haplotype TGC was the rarest observed (0.005), and additional three had very low frequency (from 0.011 to 0.042). Sudanese camels showed a higher frequency of the haplotype GAC (0.348) compared to the Nigerian (0.187), where the most

represented haplotype was GGC (0.290), rather underrepresented in the Sudanese camels (0.028). Ecotypes within Sudanese population showed further differences. For instance, Shanbali and Lahaoi *vs.* Khali and Arabi showed nearly opposite frequencies for the haplotypes GAC and GAT, thus potentially opening the possibility for a rapid directional selection if future studies will demonstrate associations with milk properties.

The knowledge of haplotypes is particular useful in breeding schemes because they may impact on a trait in a different way compared to single alleles, exploiting all the genetic effects existing among individual genes. This would be particularly convenient for the casein genes, which are closely linked. Therefore, a deeper screening of casein variability should be accomplished in dairy camels at both protein and DNA level to have a better knowledge on the amount and potential use of the genetic polymorphisms at these loci.

Cytogenetic Mapping

Casein genes are mapped on the same chromosome in all species investigated so far. For instance, they are located on chromosome 6 in cattle, sheep and goat, on chromosome 4 in humans, 8 in pig, 14 in rat, 3 in horse, etc. (Rijnkels, 2002, Martin et al., 2013). Conversely, the cytogenetic map of the casein genes was never reported in camelids and, in general, very little information is available so far on the physical mapping of other *loci* (Avila et al., 2014a; Avila et al., 2014b; Perelman et al., 2018).

The production of specific probes allowed mapping the casein genes to the chromosome 2q21. Such result also confirms the comparative evolutionary study of Balmus et al. (2007). In fact, cross-hybridization experiments with molecular painting probes evidenced that the dromedary camel chromosome 2 (CDR2) corresponds to the bovine chromosome 6 (BTA6) where the casein genes have been mapped (Rijnkels, 2002). Furthermore, the extensive similarities reported in the karyotypes of the camelids (*Camelus dromedarius*, *Camelus bactrianus*, *Lama glama*, *Lama guanicoe*, *V. pacos* and *Vicugna vicugna*) (Bunch et al., 1985; Di Berardino et al., 2006; Balmus et al., 2007) confirm that CDR2 and VPA2 are homologous chromosomes of related species.

This result is also interesting for its potential to physically map other genes on camel chromosome 2. For instance, the spotting *locus* responsible of white-spotting phenotypes in cattle was mapped on BTA6, in a chromosomal region including the *KIT* gene (Grosz and MacNeil, 1999), approximately 15 Mbp upstream the casein cluster. The white-spotting phenotype is an undesired characteristic in alpacas, which are mainly bred for the quality of their coat fibers (extremely fine, hypoallergenic, and naturally stained). Therefore, studies on the genetic variability of the casein cluster in alpacas might be of interest to identify and select alleles in linkage disequilibrium with favorable coat characteristics. On the other site, the so called "blue-eyed white phenotypes" are in some cases associated with congenital deafness (Gauly et al., 2005) and associated with the *KIT* locus in many species, including alpacas (Jackling et al., 2014).

CONCLUSION

The knowledge of casein genes in camelids herein summarized provide fundamental information useful for different applications, such as biodiversity analysis or association studies functional characteristics (dietetic, technological, and nutraceutical) of camel milk to better meet the consumers' requirements.

Nowadays, planning the production of milk with different protein properties suitable for its specific use is a realistic challenge for breeders and an important goal for animal geneticists. In this respect, all the genetic variability found is useful in selection programs of dairy camels for better exploiting the effects of the entire casein cluster on milk yield and its related traits.

AUTHOR CONTRIBUTIONS

AP and GE conceived and designed the experiments. AP performed the experiments. AP and GC analyzed the data. GE and LD contributed reagents/materials/analysis tools. AP wrote the paper. AP, ETS, MDO, GC, LD, and GE revised the article critically for important intellectual content. AP, ETS, MDO, GC, LD, and GE gave final approval of the version to be published.

FUNDING

This research was financially supported by the project Camilk (PAUA_CONTR_FIN_18_01), the King Baudouin Foundation United States (KBFUS) Grant number 20180252.

ACKNOWLEDGMENTS

The authors thank Dr. Henrik Wagner from the Department of Obstetrics, Gynaecology and Andrology of Large and Small Animals with ambulance, Justus-Liebig University of Giessen (Germany) for proving alpaca samples for karyotyping.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00748/full#supplementary-material

REFERENCES

Abdussamad, A., Charruau, P., Kalla, D., and Burger, P. (2015). Validating local knowledge on camels: colour phenotypes and genetic variation of dromedaries in the Nigeria-Niger corridor. *Lives Sci.* 181, 131–136. doi: 10.1016/j.livsci.2015.07.008

Al hay, O. A., and Al Kanhal, H. A. (2010). Compositional, technological and nutritional aspects of dromedary camel milk. *Int. Dairy J.* 20 (12), 811–821. doi: 10.1016/j.idairyj.2010.04.003

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-terWengel, P., Pitt, D., et al. (2016). Ancient and modern DNA reveal dynamics of domestication

and cross-continental dispersal of the dromedary. *PNAS* 113 (24), 6707–6712. doi: 10.1073/pnas.1519508113

Avila, F., Baily, M. P., Perelman, P., Das, P. J., Pontius, J., Chowdhary, R., et al. (2014a). A comprehensive whole-genome integrated cytogenetic map for the alpaca (*Lama pacos*). *Cytogenet. Genome Res.* 144, 196–207. doi: 10.1159/000370329

Avila, F., Das, P. J., Kutzler, M., Owens, E., Perelman, P., Rubes, J., et al. (2014b). Development and application of camelid molecular cytogenetic tools. *J. Hered* 105 (6), 858–869. doi: 10.1093/jhered/ess067

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C., Alkalaeva, E. S., Fu, B., et al. (2007). Cross-species chromosome painting among camel, cattle, pig and human: further insights into the putative Cetartiodactyla ancestral karyotype. *Chromosome Res.* 15 (4), 499–514. doi: 10.1007/s10577-007-1154-x

Boisnard, M., Hue, D., Bouniol, C., Mercier, J. C., and Gaye, P. (1991). Multiple mRNA species code for two non-allelic forms of ovine αs2-casein. *Eur. J. Biochem.* 201 (3), 633–641. doi: 10.1111/j.1432-1033.1991.tb16324.x

Bouniol, C., Printz, C., and Mercier, J. C. (1993). Bovine αs2-casein D is generated by exon VIII skipping. Gene 128 (2), 289–293. doi: 10.1016/0378-1119(93)90577-P

Bowen, N. J., and Jordan, I. K. (2002). Transposable elements and the evolution of eukaryotic complexity. *Curr. Issues Mol. Bio.* 4, 65–76. doi.org/10.21775/cimb.004.065

Bunch, T. D., Foote, W. C., and Maciulis, A. (1985). Chromosome banding pattern homologies and NORs for the Bactrian camel, guanaco and llama. *J. Hered* 76, 115–118. doi: 10.1093/oxfordjournals.jhered.a110034

Carneiro, M., and Ferrand, N. (2007). Extensive intragenic recombination and patterns of linkage disequilibrium at the *CSN3* locus in European rabbit. *Genet. Sel. Evol.* 39 (3), 341. doi: 10.1051/gse:2007007

Caroli, A., Chiatti, F., Chessa, S., Rignanese, D., Bolla, P., and Pagnacco, G. (2006). Focusing on the goat casein complex. *J. Dairy Sci.* 89 (8), 3178–3187. doi: 10.3168/jds.S0022-0302(06)72592-9

Caroli, A. M., Chessa, S., and Erhardt, G. J. (2009). Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. *J. Dairy Sci.* 92 (11), 5335–5352. doi: 10.3168/jds.2009-2461

Christoffels, V. M., Grange, T., Kaestner, K. H., Cole, T. J., Darlington, G. J., Croniger, C. M., et al. (1998). Glucocorticoid receptor, C/EBP, HNF3, and protein kinase A coordinately activate the glucocorticoid response unit of the carbamoylphosphate synthetase I gene. *Mol. Cell. Biol.* 18 (11), 6305–6315. doi: 10.1128/MCB.18.11.6305

Cosenza, G., Pauciullo, A., Colimoro, L., Mancusi, A., Rando, A., Di Berardino, D., et al. (2007). A SNP in the goat CSN2 promoter region is associated with the absence of β-casein in milk. *Anim. Genet.* 38 (6), 655–658. doi: 10.1111/j.1365-2052.2007.01649.x

Cosenza, G., Pauciullo, A., Gallo, D., Colimoro, L., D'avino, A., Mancusi, A., et al. (2008). Genotyping at the *CSN1S1* locus by PCR-RFLP and AS-PCR in a Neapolitan goat population. *Small Ruminant Res.* 74 (1–3), 84–90. doi: 10.1016/j.smallrumres.2007.03.010

Cosenza, G., Pauciullo, A., Feligini, M., Coletta, A., Colimoro, L., Di Berardino, D., et al. (2009). A point mutation in the splice donor site of intron 7 in the αs2-casein encoding gene of the Mediterranean River buffalo results in an allele-specific exon skipping. *Anim. Genet.* 40 (5), 791. doi: 10.1111/j.1365-2052.2009.01897.x

Cosenza, G., Pauciullo, A., Annunziata, A. L., Rando, A., Chianese, L., Marletta, D., et al. (2010). Identification and characterization of the donkey CSN1S2 I and II cDNAs. *Ital. J. Anim. Sci.* 9 (2), e40. doi: 10.4081/ijas.2010.e40

Di Berardino, D., Nicodemo, D., Coppola, G., King, A., Ramunno, L., Cosenza, G., et al. (2006). Cytogenetic characterization of alpaca (*Lama pacos*, fam. Camelidae) prometaphase chromosomes. *Cytogenet. Genome Res.* 115 (2), 138–144. doi: 10.1159/000095234

Erhardt, G., Lisson, M., Weimann, C., Wang, Z., El Zubeir, I. E. Y. M., and Pauciullo, A. (2016). Alpha S1-casein polymorphisms in camel (*Camelus dromedarius*). *Trop. Anim. Health Prod.* 48 (5), 879–887. doi: 10.1007/s11250-016-0997-6

Erhardt, G., Gu, M., Wagner, H., Di Stasio, L., and Pauciullo, A., (2017). *Vicugna pacos* αs1-casein: identification of new polymorphisms at the *CSN1S1* gene. *Proceedings of the 7th European Symposium on South American Camelids and 3rd European Meeting on Fibre Animals*; June, 12–17; Italy: Assisi, 36.

Evdotchenko, D., Han, Y., Bartenschlager, H., Preuss, S., and Geldermann, H. (2003). New polymorphic microsatellite loci for different camel species. *Mol. Ecol. Notes* 3 (3), 431–434. doi: 10.1046/j.1471-8286.2003.00477.x

Farah, Z., Mollet, M., Younan, M., and Dahir, R. (2007). Camel dairy in Somalia: limiting factors and development potential. *Livest Sci.* 110 (1–2), 187–191. doi: 10.1016/j.livsci.2006.12.010

Faye, B., Grech, S., and Korchani, T. (2002). Le dromadaire, entre féralisation et intensification. *Anthropozoos* 39 (2), 7–13.

Faye, B., and Konuspayeva, G. (2012). The sustainability challenge to the dairy sector—the growing importance of non-cattle milk production worldwide. *Int. Dairy J.* 24 (2), 50–56. doi: 10.1016/j.idairyj.2011.12.011

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2016). The *de novo* genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16 (1), 314–324. doi: 10.1111/1755-0998.12443

Fujiwara, Y., Miwa, M., Nogami, M., Okumura, K., Nobori, T., Suzuki, T., et al. (1997). Genomic organization and chromosomal localization of the human casein gene family. *Hum. Genet.* 99 (3), 368–373. doi: 10.1007/s004390050374

Gauly, M., Vaughan, J., Hogreve, S. K., and Erhardt, G. (2005). Brainstem auditory-evoked potential assessment of auditory function and congenital deafness in llamas (*Lama glama*) and alpacas (*L pacos*). *J. Vet. Intern. Med.* 19 (5), 756–760. doi: 10.1111/j.1939-1676.2005.tb02757.x

Giambra, I. J., Chianese, L., Ferranti, P., and Erhardt, G. (2010). Genomics and proteomics of deleted ovine CSN1S1∗ I. *Int. Dairy J.* 20 (3), 195–202. doi: 10.1016/j.idairyj.2009.09.005

Ginger, M. R., Piotte, C. P., Otter, D. E., and Grigor, M. R. (1999). Identification, characterisation and cDNA cloning of two caseins from the common brushtail possum (*Trichosurus vulpecula*) 1. *Biochim. Biophys. Acta Gen. Subj.* 1427 (1), 92–104. doi: 10.1016/S0304-4165(99)00008-2

Grosz, M. D., and MacNeil, M. D. (1999). The 'spotted' locus maps to bovine chromosome 6 in a Hereford-cross population. *J. Hered.* 90 (1), 233–236. doi: 10.1093/jhered/90.1.233

Harvatine, K. J., and Bauman, D. E. (2006). SREBP1 and thyroid hormone responsive spot 14 (S14) are involved in the regulation of bovine mammary lipid synthesis during diet-induced milk fat depression and treatment with CLA. *J. Nutr.* 136 (10), 2468–2474. doi: 10.1093/jn/136.10.2468

Hobor, S., Kunej, T., and Dovc, P. (2008). Polymorphisms in the kappa casein (*CSN3*) gene in horse and comparative analysis of its promoter and coding region. *Anim. Genet.* 39 (5), 520–530. doi: 10.1111/j.1365-2052.2008.01764.x

Iannuzzi, L., and Di Berardino, D. (2008). Tools of the trade: diagnostics and research in domestic animal cytogenetics. *J. Appl. Genet.* 49 (4), 357–366. doi: 10.1007/BF03195634

Inman, C. K., Li, N., and Shore, P. (2005). Oct-1 counteracts autoinhibition of Runx2 DNA binding to form a novel Runx2/Oct-1 complex on the promoter of the mammary gland-specific gene β-casein. *Mol. Cell. Biol.* 25 (8), 3182–3193. doi: 10.1128/MCB.25.8.3182-3193.2005

Jackling, F. C., Johnson, W. E., and Appleton, B. R. (2014). The genetic inheritance of the blue-eyed white phenotype in alpacas (*Vicugna pacos*). *J. Hered* 105 (6), 941–951. doi: 10.1093/jhered/ess093

Jiang, Z., and Zacksenhaus, E. (2002). Activation of retinoblastoma protein in mammary gland leads to ductal growth suppression, precocious differentiation, and adenocarcinoma. *J. Cell. Biol.* 156 (1), 185–198. doi: 10.1083/jcb.200106084

Jones, W. K., Yu-Lee, L., Clift, S. M., Brown, T. L., and Rosen, J. (1985). The rat casein multigene family. Fine structure and evolution of the beta-casein gene. *J. Biol. Chem.* 260 (11), 7042–7050.

Kappeler, S., Farah, Z., and Puhan, Z. (1998). Sequence analysis of *Camelus dromedarius* milk caseins. *J. Dairy Res.* 65 (2), 209–222. doi: 10.1017/S0022029997002847

Kappeler, S., Farah, Z., and Puhan, Z. (2003). 5′-Flanking regions of camel milk genes are highly similar to homologue regions of other species and can be divided into two distinct groups. *J. Dairy Sci.* 86 (2), 498–508. doi: 10.3168/jds.S0022-0302(03)73628-5

Kawasaki, K., Lafont, A. G., and Sire, J. Y. (2011). The evolution of milk casein genes from tooth genes before the origin of mammals. *Mol. Biol. Evol.* 28, 2053–2061. doi: 10.1093/molbev/msr020

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: repbasesubmitter and censor. *BMC bioinf.* 7 (1), 474. doi: 10.1186/1471-2105-7-474

Konuspayeva, G., Faye, B., and Loiseau, G. (2009). The composition of camel milk: a meta-analysis of the literature data. *J. Food Compos. Anal.* 22, 95–101. doi: 10.1016/j.jfca.2008.09.008

Leyva, V., and Markas, J. (1991). Involucion de la glandula mamaria en alpacas y efecto sobre el peso corporal y produccion de fibra. *Turrialba* 41, 59–63.

Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165 (4), 2213–2233.

Liu, G., Zhao, S., Bailey, J. A., Sahinalp, S. C., Alkan, C., Tuzun, E., et al. (2003). Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* 13 (3), 358–368. doi: 10.1101/gr.923303

Martin, P., Cebo, C., and Miranda, G. (2013). "Interspecies comparison of milk proteins: quantitative variability and molecular diversity," in *Advanced Dairy Chemistry: Vol 1A Proteins: Basic Aspects*, 4th; Boston, MA, USA: Springer. 387–429. doi: 10.1007/978-1-4614-4714-6_13

Mburu, D., Ochieng, J., Kuria, S., Jianlin, H., Kaufmann, B., Rege, J., et al. (2003). Genetic diversity and relationships of indigenous Kenyan camel (*Camelus dromedarius*) populations: implications for their classification. *Anim. Genet.* 34 (1), 26–32. doi: 10.1046/j.1365-2052.2003.00937.x

McKnight, T. L. (1969). *The camel in Australia*. Carlton Vict. Australia: Melbourne University Press.

Menon, R. S., Chang, Y.-F., Jeffers, K. F., and Ham, R. G. (1992). Exon skipping in human β-casein. *Genomics* 12 (1), 13–17. doi: 10.1016/0888-7543(92)90400-M

Milenkovic, D., Martin, P., Guérin, G., and Leroux, C. (2002). A specific pattern of splicing for the horse α S1-Casein mRNA and partial genomic characterization of the relevant locus. *Genet. Sel. Evol.* 34 (4), 509. doi: 10.1186/1297-9686-34-4-509

Morin, D. E., and Rowan, L. L. (1995). Composition of milk from llamas in the United States. *J. Dairy Sci.* 78, 1713–1720. doi: 10.3168/jds.S0022-0302(95)76796-0

Nagy, P., Thomas, S., Markó, O., and Juhász, J. (2012). Milk production, raw milk quality and fertility of dromedary camels (*Camelus dromedarius*) under intensive management. *Acta Vet. Hung.* 61 (1), 71–84. doi: 10.1556/AVet.2012.051

Nijman, I. J., van Tessel, P., and Lenstra, J. A. (2002). SINE retrotransposition during the evolution of the Pecoran ruminants. *J. Mol. Evol.* 54 (1), 9–16. doi: 10.1007/s00239-001-0012-2

Pauciullo, A., Shuiep, E., Cosenza, G., Ramunno, L., and Erhardt, G. (2013a). Molecular characterization and genetic variability at κ-casein gene (*CSN3*) in camels. *Gene* 513 (1), 22–30. doi: 10.1016/j.gene.2012.10.083

Pauciullo, A., Fleck, K., Lühken, G., Di Berardino, D., and Erhardt, G. (2013b). Dual-color high-resolution fiber-FISH analysis on lethal white syndrome carriers in sheep. *Cytogenet. Genome Res.* 140 (1), 46–54. doi: 10.1159/000350786

Pauciullo, A., Giambra, I., Iannuzzi, L., and Erhardt, G. (2014). The β-casein in camels: molecular characterization of the *CSN2* gene, promoter analysis and genetic variability. *Gene* 547 (1), 159–168. doi: 10.1016/j.gene.2014.06.055

Pauciullo, A., and Erhardt, G. (2015). Molecular characterization of the llamas (Lama glama) casein cluster genes transcripts (*CSN1S1, CSN2, CSN1S2, CSN3*) and regulatory regions. *PloS one* 10 (4), e0124963. doi: 10.1371/journal.pone.0124963

Pauciullo, A., Knorr, C., Perucatti, A., Iannuzzi, A., Iannuzzi, L., and Erhardt, G. (2016). Characterization of a very rare case of living ewe-buck hybrid using classical and molecular cytogenetics. *Sci. Rep.* 6, 34781. doi: 10.1038/srep34781

Pauciullo, A., Gauly, M., Cosenza, G., Wagner, H., and Erhardt, G. (2017). Lama glama αs1-casein: identification of new polymorphisms in the *CSN1S1* gene. *J. Dairy Sci.* 100 (2), 1282–1289. doi: 10.3168/jds.2016-11918

Pauloin, A., Rogel-Gaillard, C., Piumi, F., Hayes, H., Fontaine, M. L., Chanat, E., et al. (2002). Structure of the rabbit αs1-and β-casein gene cluster, assignment to chromosome 15 and expression of the αs1-casein gene in HC11 cells. *Gene* 283 (1), 155–162. doi: 10.1016/S0378-1119(01)00872-1

Penedo, M., Caetano, A., and Cordova, K. (1999). Eight microsatellite markers for South American camelids. *Anim. Genet.* 30 (2), 166–167. doi: 10.1046/j.1365-2052.1999.00382-8.x

Perelman, P. L., Pichler, R., Gaggl, A., Larkin, D. M., Raudsepp, T., Alshanbari, F., et al. (2018). Construction of two whole genome radiation hybrid panels for dromedary (*Camelus dromedarius*): 5000 RAD and 15000 RAD. *Sci. Rep.* 8, 1982. doi: 10.1038/s41598-018-20223-5

Pérez, M. J., Leroux, C., Bonastre, A. S., and Martin, P. (1994). Occurrence of a LINE sequence in the 3′ UTR of the goat αs1-casein E-encoding allele associated with reduced protein synthesis level. *Gene* 147 (2), 179–187. doi: 10.1016/0378-1119(94)90063-9

Ramunno, L., Cosenza, G., Rando, A., Illario, R., Gallo, D., Di Berardino, D., et al. (2004). The goat αs1-casein gene: gene structure and promoter analysis. *Gene* 334, 105–111. doi: 10.1016/j.gene.2004.03.006

Ramunno, L., Cosenza, G., Rando, A., Pauciullo, A., Illario, R., Gallo, D., et al. (2005). Comparative analysis of gene sequence of goat CSN1S1 F and N alleles and characterization of *CSN1S1* transcript variants in mammary gland. *Gene* 345 (2), 289–299. doi: 10.1016/j.gene.2004.12.003

Rando, A., Di Gregorio, P., Ramunno, L., Mariani, P., Fiorella, A., Senese, C., et al. (1998). Characterization of the *CSN1A*G Allele of the Bovine αs1-Casein locus by the insertion of a relict of a long interspersed element. *J. Dairy Sci.* 81 (6), 1735–1742. doi: 10.3168/jds.S0022-0302(98)75741-8

Reed, B. D., Charos, A. E., Szekely, A. M., Weissman, S. M., and Snyder, M. (2008). Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes. *PLoS Genet.* 4 (7), e1000133. doi: 10.1371/journal.pgen.1000133

Riek, A., and Gerken, M. (2006). Changes in llama (*Lama glama*) milk composition during lactation. *J. Dairy Sci.* 89, 3484–3493. doi: 10.3168/jds.S0022-0302(06)72387-6

Rijnkels, M. (2002). Multispecies comparison of the casein gene loci and evolution of casein gene family. *J. Mammary Gland Biol. Neoplasia* 7 (3), 327–345. doi: 10.1023/A:1022808918013

Rijnkels, M., Wheeler, D., De Boer, H., and Pieper, F. (1997). Structure and expression of the mouse casein gene locus. *Mamm. Genome* 8 (1), 9–15. doi: 10.1007/s003359900338

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin, P. (2018). Combining different proteomic approaches to resolve complexity of the milk protein fraction of dromedary, Bactrian camels and hybrids, from different regions of Kazakhstan. *PloS one* 13 (5), e0197026. doi: 10.1371/journal.pone.0197026

Ryskaliyeva, A. (2018). Exploring the fine composition of Camelus milk from Kazakhstan with emphasis on protein components. PhD Thesis. CIRAD INRA. https://umr-selmet.cirad.fr/en/news/soutenance-de-these-d-alma-ryskaliyeva.

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin, P. (2019). Alternative splicing events expand molecular diversity of camel CSN1S2 increasing its ability to generate potentially bioactive peptides. *Sci. Rep.* 9, 5243. doi: 10.1038/s41598-019-41649-5

Saadaoui, B., Bianchi, L., Henry, C., Miranda, G., Martin, P., and Cebo, C. (2014). Combining proteomic tools to characterize the protein fraction of llama (*Lama glama*) milk. *Electrophoresis* 35, 1406–1418. doi: 10.1002/elps.201300383

Sambrook, J., Fritsch, E. F., and Maniatis, T., (1989). *Molecular cloning*. New York, NY: Cold Spring Harbor.

Schild, T., and Geldermann, H. (1996). Variants within the 5′ -flanking regions of bovine milk-protein-encoding genes. III. Genes encoding the Ca-sensitive caseins αs1, αs2 and β. *Theor. Appl. Genet.* 93 (5–6), 887–893. doi: 10.1007/BF00224090

Shuiep, E., Giambra, I. J., El Zubeir, I. E. Y. M., and Erhardt, G. (2013). Biochemical and molecular characterization of polymorphisms of α s1-casein in Sudanese camel (*Camelus dromedarius*) milk. *Int. Dairy J.* 28 (2), 88–93. doi: 10.1016/j.idairyj.2012.09.002

Spencer, P., Wilson, K., and Tinson, A. (2010). Parentage testing of racing camels (*Camelus dromedarius*) using microsatellite DNA typing. *Anim. Genet.* 41 (6), 662–665. doi: 10.1111/j.1365-2052.2010.02044.x

Stewart, A. F., Bonsing, J., Beattie, C. W., Shah, F., Willis, I. M., and Mackinlay, A. G. (1987). Complete nucleotide sequences of bovine alpha S2-and beta-casein cDNAs: comparisons with related sequences in other species. *Mol. Biol. Evol.* 4 (3), 231–241. doi: 10.1093/oxfordjournals.molbev.a040437

Subramaniam, N., Campión, J., Rafter, I., and Okret, S. (2003). Cross-talk between glucocorticoid and retinoic acid signals involving glucocorticoid receptor interaction with the homoeodomain protein Pbx1. *Biochem. J.* 370 (3), 1087–1095. doi: 10.1042/bj20020471

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24 (8), 1596–1599. doi: 10.1093/molbev/msm092

Wang, Z., Ding, G., Chen, G., Sun, Y., Sun, Z., Zhang, H., et al. (2012). Genome sequences of wild and domestic bactrian camels. *Nat. Commun.* 3, 1202–1202. doi: 10.1038/ncomms2192

Weimann, C., Meisel, H., and Erhardt, G. (2009). Bovine κ-casein variants result in different angiotensin I converting enzyme (ACE) inhibitory peptides. *J. Dairy Sci.* 92 (5), 1885–1888. doi: 10.3168/jds.2008-1671

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5, 5188. doi: 10.1038/ncomms6188

Wyszomierski, S. L., and Rosen, J. M. (2001). Cooperative effects of STAT5 (signal transducer and activator of transcription 5) and C/EBP β (CCAAT/enhancer-binding protein-β) on β-casein gene transcription are mediated by the glucocorticoid receptor. *Mol. Endocrinol.* 15 (2), 228–240. doi: 10.1210/mend.15.2.0597

Zhao, F. Q., Adachi, K., and Oka, T. (2002). Involvement of Oct-1 in transcriptional regulation of β-casein gene expression in mouse mammary gland. *Biochim. Biophys. Acta Gene Struct. Expression* 1577 (1), 27–37. doi: 10.1016/S0167-4781(02)00402-5

Zhao, D., Bai, Y., and Niu, Y. (2015). Composition and characteristics of Chinese Bactrian camel milk. *Small Ruminant Res.* 127, 58–67. doi: 10.1016/j.smallrumres.2015.04.008

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Pauciullo, Shuiep, Ogah, Cosenza, Di Stasio and Erhardt. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 September 2019
doi: 10.3389/fgene.2019.00893

Genome Diversity and Signatures of Selection for Production and Performance Traits in Dromedary Camels

Hussain Bahbahani[1], Hassan H. Musa[2]*, David Wragg[3], Eltahir S. Shuiep[4], Faisal Almathen[5] and Olivier Hanotte[6]*

[1] Department of Biological Sciences, Faculty of Science, Kuwait University, Kuwait City, Kuwait, [2] Department of Medical Microbiology, Faculty of Medical Laboratory Sciences, University of Khartoum, Khartoum North, Sudan, [3] Centre for Tropical Livestock Genetics and Health, The Roslin Institute, Edinburgh, United Kingdom, [4] Department of Animal Production, Faculty of Agricultural and Environmental Sciences, University of Gadarif, Gadarif State, Sudan, [5] Department of Public Health, College of Veterinary Medicine, King Faisal University, Al-Hasa, Saudi Arabia, [6] LiveGene, International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia

OPEN ACCESS

Edited by:
Stéphane Joost, École Polytechnique Fédérale de Lausanne, Switzerland

Reviewed by:
Filippo Biscarini, Italian National Research Council (CNR), Italy
Pablo Orozco-terWengel, Cardiff University, United Kingdom

Correspondence:
Hussain Bahbahani
hussain.bahbahani@ku.edu.kw;
hussain.bahbahani@gmail.com
Hassan H. Musa
hassanhm@uofk.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *17 January 2019*
Accepted: *23 August 2019*
Published: *19 September 2019*

Citation:
Bahbahani H, Musa HH, Wragg D, Shuiep ES, Almathen F and Hanotte O (2019) Genome Diversity and Signatures of Selection for Production and Performance Traits in Dromedary Camels. Front. Genet. 10:893. doi: 10.3389/fgene.2019.00893

Dromedary camels (*Camelus dromedarius*) are single-humped animals found throughout the deserts of Africa, the Arabian Peninsula, and the southwest of Asia. This well-adapted species is mainly used for milk and meat production, although some specific types exhibit superior running performance and are used in racing competitions. However, neither performance nor production camels are bred under intensive genomic selection programs with specific aims to improve these traits. In this study, the full genome sequence data of six camels from the Arabian Peninsula and the genotyping-by-sequencing data of 44 camels (29 packing and 15 racing) from Sudan were analyzed to assess their genome diversities, relationships, and candidate signatures of positive selection. Genome ADMIXTURE and principle component analyses indicate clear geographic separation between the Sudanese and the Arabian Peninsula camels, but with no population-specific genetic distinction within populations. Camel samples from the Arabian Peninsula show higher mean heterozygosity (0.560 ± 0.003) than those from Sudan (0.347 ± 0.003). Analyses of signatures of selection, using pooled heterozygosity (*Hp*) approach, in the Sudanese camels revealed 176, 189, and 308 candidate regions under positive selection in the combined and packing and racing camel populations, respectively. These regions host genes that might be associated with adaptation to arid environment, dairy traits, energy homeostasis, and chondrogenesis. Eight regions show high genetic differentiation, based on *Fst* analysis, between the Sudanese packing and racing camel types. Genes associated with chondrogenesis, energy balance, and urinary system development were found within these regions. Our results advocate for further detailed investigation of the genome of the dromedary camel to identify and characterize genes and variants associated with their valuable phenotypic traits. The results of which may support the development of breeding programs to improve the production and performance traits of this unique domesticated species.

Keywords: genotype-by-sequence, positive selection, milk production, racing camels, Arabian Peninsula, Sudan

INTRODUCTION

The *Camelidae* family is divided into two tribes, the New World camel (Lamini) and the Old World camel (Camelini). Within the Camelini, there are two domesticated species, the two-humped Bactrian camel (*Camelus bactrianus*) and the single-humped dromedary camel (*Camelus dromedarius*), in addition to the critically endangered two-humped wild Bactrian camel (*Camelus ferus*) found in northern China and southern Mongolia (Wilson, 1998).

Out of approximately 35 million camel heads worldwide (FAO, 2019), the majority (95%) are of the dromedary type (Hashim et al., 2015). Unlike the Bactrian camels, which are distributed throughout central and eastern Asia, dromedary camels mainly populate the desert and semi-desert areas across Africa, the Arabian Peninsula, and southwest of Asia (Wilson, 1998). They are highly adapted to the harsh desert environment, which is characterized by high temperatures and scarcity of food and water. Dromedary camels are tolerant to temperatures in excess of 40°C and can survive for up to 20 to 35 days without water, losing up to 25% of their body weight (Schmidt-Nielsen, 1959; Musa et al., 2006).

Archaeological evidence indicates dromedary camels to have been domesticated in the southeast of the Arabian Peninsula in the late second millennium BC (Magee, 2015), following which they spread to the northern part of the Arabian Peninsula and Africa *via* ancient trading routes. Being highly adapted to the desert environment, dromedary camels were historically the preferred means of transporting people and goods along routes such as the "incense road"—the trans-Arabian trading route used to transport valuable spices and perfumes from southern to northern areas of the Arabian Peninsula. The connection between the Arabian Peninsula and Africa was first established in East of Africa, through the Islands of Socotra, in association with sea-borne incense trading during the 1st millennium BC (Epstein and Mason, 1971). In parallel, luxury goods, such as ivory, wools, and skins, were also transported from Africa to the Arabian Peninsula through Aden (Andree-Salvini, 2010). By the 7th century, dromedary camels were widespread throughout the African Sahara employed in the transport of goods (Gauthier-Pilters and Anne, 1983). The spread of Islam to Africa at that time also contributed to the spread of dromedary between Africa and the Arabian Peninsula, as they were used to transport people to Makkah in the Arabian Peninsula during the annual pilgrimage "*hajj*" (Wilson, 1998).

The majority of dromedary, around 80%, are found in Africa. Sudan accommodates around 4.8 million heads, second in size to Somalia (Ali et al., 2019), representing around 22% of the animal biomass in the country (Eisa and Mustafa, 2011). The history of dromedary camel in Sudan can be traced to 15–25 years BC following human migrations from Egypt (Eisa and Mustafa, 2011; Hashim et al., 2015). Camels inhabit an area in Sudan characterized by arid conditions and erratic periods of rainfall. The area is flanked by the Butana plains and Red Sea hills in the East, and the Darfur and Kordofan States in the West (Eisa and Mustafa, 2011). Here, camels are classified as either "packing" or "racing" with the former type being used for milk and meat production, while the latter is typically used in racing competitions. Packing camel populations, such as the Arabi, Kenanai, Lahawi, and Rashaydi, constitute around 80% of a nomad's herd and are characterized by their heavy weight and for having a well-developed hump (Hashim et al., 2015). Milk production in these camels is in the range of 820–2,400 liters per lactation period, which typically lasts for 12–18 months. The level of productivity depends mainly on the season and the farming management system (Faye et al., 2011). Other camel populations in Sudan, such as the Anafi and Bishari, which are famous for their racing performance, are lighter than the pack camels and are generally found in the northeast of Sudan and the River Nile State (Wardeh, 2004; Hashim et al., 2015). Populations of each of these camel types in Sudan are named after the ethnic groups that rear them.

Dromedaries in the Arabian Peninsula follow a similar classification system to that in Africa (Wardeh, 2004) but are further classified according to their coat color. Black dromedary camels are called Majaheem, whereas white and brown camels are called Wodh and Shual, respectively. A further type also characterized by a brown coat, but darker than Shual on the hump and tail, is called Sofor (Al-Swailem et al., 2007; Almathen et al., 2018). Each of these types are considered to be packing animals used for meat and milk production, with the Majaheem type being the most popular (Wardeh, 2004). Other popular types include the Omani originating from Oman, and the Hura from Saudi Arabia, which are both considered high-performance racing camels (Wardeh, 2004). Despite the different classifications and livestock functions, no selection programs aimed at improving the different types are in place.

Until recently, genetic analyses of dromedary camel have been largely restricted to studies involving autosomal microsatellite markers (Mburu et al., 2003; Almathen et al., 2016), partial mitochondrial DNA (mtDNA) sequences (Babar et al., 2015; Almathen et al., 2016) and candidate gene sequencing (Pauciullo et al., 2013; Shuiep et al., 2013; Almathen et al., 2018). These tools have mainly been employed to assess the genetic diversity and structure of dromedary camel populations from different geographical locations, e.g., Kenya (Mburu et al., 2003), the Arabian Peninsula (Almathen et al., 2016), and Pakistan (Babar et al., 2015). Efforts to link genotypes to phenotypes have also been undertaken—for example, Almathen et al. (2018) linked an arginine to cysteine missense mutation at protein position 301 in the melanocortin 1 receptor (*MC1R*) gene to white coat color. In the same study, a 1-bp deletion (23del T/T) and a SNP (25G/A) in exon 2 of the agouti signaling protein (*ASIP*) gene were linked to the black and dark-brown coat color phenotypes of Saudi Arabian dromedary. Recently, Khalkhali-Evrigh et al. (2018) analyzed the full genomes of two Iranian dromedary camels and reported non-synonymous variants in genes related to adaptation to the desert environment.

Signatures of selection analyses in African indigenous livestock have been conducted in different breeds or populations for a number of species, including, for instance, cattle (Bahbahani et al., 2015; Bahbahani et al., 2017; Kim et al., 2017a; Bahbahani et al., 2018a; Bahbahani et al., 2018b), goat (Kim et al., 2016; Onzima et al., 2018), and sheep (Kim et al., 2016), but to date,

nothing has been reported for dromedary. Reference genomes of dromedary camels from the Arabian Peninsula and North Africa were published in 2014 and 2015, respectively (Wu et al., 2014; Fitak et al., 2015); however, neither is very contiguous, being assembled into 32,573 and 35,752 scaffolds, respectively. These assemblies are a fundamental first-step toward facilitating genome-wide analyses for signatures of selection. For example, Wu et al. (2014) analysis of the Arabian Peninsula dromedary genome revealed accelerated evolution and positive selection in genes related to fat metabolism, heat stress response, and salt metabolism.

Although the costs of whole-genome sequencing (WGS) have declined significantly in recent years, it remains an expensive endeavour to sequence large numbers of individuals, with large and complex genomes, to enable genome-wide studies. For some livestock species such as cattle and sheep, high-density genotyping microarrays are available enabling hundreds of thousands of genomic positions to be genotyped at a fraction of the cost of WGS (Rincon et al., 2011; Kijas et al., 2014). Where these tools are unavailable, typically in the case of non-model species with poor genomic resources, an alternative approach is to employ a genotyping-by-sequencing (GBS) strategy. GBS is a high-throughput multiplex sequencing system based on constructing reduced representative libraries for subsequent sequencing (Elshire et al., 2011). As with genotype microarrays, GBS can be performed at a fraction of the cost of WGS and does not suffer the ascertainment bias associated with array-based genotyping (Poland and Rife, 2012). The main disadvantage of this approach, however, is that the distribution of coverage throughout the genome is generally not uniform, and as a result, there is typically extensive between-sample variation in the genomic regions sequenced (Beissinger et al., 2013). GBS has been widely used in plant breeding for genome-wide association analysis, genomic diversity studies, and genomic selection (reviewed in He et al. (2014)). In addition, Wang et al. (2018) employed GBS to analyze the genomes of Chinese Landrace and Yorkshire pigs, identifying candidate signatures of selection in genes related to fatty acid biosynthesis, animal growth and development, and immune responses.

The availability of reference genome assemblies coupled with GBS enables us to explore in greater detail than ever before the genomes of different dromedary populations and to seek to identify genetic associations with different phenotypic traits. The aim of this study is to exploit these resources to assess the genomic diversity and relationship of dromedary camel populations from Sudan and the Arabian Peninsula and to undertake signatures of selection analyses to identify genes that might be associated with environmental adaptation, milk production, and racing performance in the Sudanese camels.

MATERIALS AND METHODS

Animal Resources

Five Sudanese indigenous camel populations were sampled for this study. These included three packing camel populations: Arabi from western Sudan (n = 9), Kenani from central Sudan (n = 10), and Lahawi from eastern Sudan (n = 10) and two racing camel populations from eastern Sudan: Bishari (n = 7) and Anafi (n = 8) (**Table 1** and **Supplementary Figure S1**). Six camel samples from the Arabian Peninsula were also included in this study: three Majaheem, two Omani, and one Sofor. Two Majaheem samples were collected from the Conservation and Genetic Improvement Centre at Al-Kharj in Saudi Arabia, while one Majaheem and one Sofor were sampled from Alwatani camel dairy farm in Saudi Arabia. The Omani samples were from the King Fahad camel herd (**Table 1**).

TABLE 1 | Sampling summary for the different Sudanese and Arabian Peninsula camel populations included in this study.

Population	Number of samples	Type	Location	Region	Type of data[1]	Type of analysis
Bishari	7	Racing camel	Kassala State	Eastern Sudan	GBS	Genetic diversity and selection analyses
Anafi	8	Racing camel	Kassala State	Eastern Sudan	GBS	Genetic diversity and selection analyses
Lahawi	10	Packing camel	Kassala State	Eastern Sudan	GBS	Genetic diversity and selection analyses
Kenani	10	Packing camel	Elgazira State	Central Sudan	GBS	Genetic diversity and selection analyses
Arabif	9	Packing camel	Darfur State	Western Sudan	GBS	Genetic diversity and selection analyses
Majaheem	2	Packing camel	the Conservation and Genetic Improvement Center at Al-Kharj	Saudi Arabia	WGS	Genetic diversity analyses
Majaheem	1	Packing camel	Alwatania camel dairy farm	Saudi Arabia	WGS	Genetic diversity analyses
Sofor	1	Packing camel	Alwatania camel dairy farm	Saudi Arabia	WGS	Genetic diversity analyses
Omani	2	Racing camel	King Fahad's camel herd	Saudi Arabia	WGS	Genetic diversity analyses

[1] *GBS, genotyping-by-sequencing; WGS, whole-genome sequencing.*

Genotyping-by-Sequencing (GBS) Data

Ten milliliters of whole blood were collected from each Sudanese sample using EDTA VACUETTE® tubes. Genomic DNA was extracted from these whole-blood samples using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol. The quality and quantity of the DNA were evaluated using a NanoDrop Spectrophotometer (NanoDrop Technologies, USA) and by gel electrophoresis. The extracted DNA samples were genotyped using pair-end GBS technology incorporating two restriction enzymes, *MseI* and *EcoRI*, sequenced on the Illumina HiSeq 2000 platform by Novogene (Novogene Co., Ltd., Tianjin, China). Novogene trimmed adapters from the sequence reads and discarded raw read pairs if (1) they were contaminated with adapter sequences, (2) uncertain nucleotides constituted more than 10% of either read, or (3) if low-quality nucleotides (base $Q_{phred} \leq 5$) constituted more than 50% of either read.

Whole-Genome Sequence (WGS) Data

Genomic DNA of the six Arabian Peninsula dromedary samples was extracted from 5 ml whole blood from each sample using the DNeasy® Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol. The extracted genomic DNA was sequenced using pair-end libraries on the Illumina HiSeq 2000 platform (Beijing Genomics Institute (BGI), China). BGI trimmed adapters from the sequence reads and discarded raw read pairs if low-quality nucleotides (base $Q_{phred} \leq 5$) constituted more than 50% of either read.

Processing of Sequence Data

The clean raw sequence reads of both the GBS and WGS data were mapped to the Arabian dromedary camel reference genome assembly (GCF_000767585.1) (Wu et al., 2014) using the *BWA-MEM* algorithm of Burrows–Wheeler Aligner (BWA) version 0.7.5a (Li and Durbin, 2010). This aligner employs local alignment which results in the ends of reads being soft-clipped if they fail to map sufficiently well. Picard tools version 1.119 (http://broadinstitute.github.io/picard/index.html) was used to sort the reads by coordinate and mark PCR duplicates. Marking duplicate reads results in those reads being ignored in downstream analyses both by the Genome Analysis Toolkit GATK version 4.1 (Mckenna et al., 2010) and SAMtools version 1.3.1 (Li, 2011). Mapped reads with insertions or deletions (indels) were realigned using GATK's IndelRealigner.

Single-nucleotide polymorphisms (SNPs) were called across the WGS (Arabian Peninsula) and GBS (Sudan) data, separately, using GATK's HaplotypeCaller and SAMtools mpileup. HaplotypeCaller employs a number of default read filters; these include: removing reads that fail platform/vendor checks, unmapped reads, secondary alignments, reads that are not well-formed, reads that fail the minimum mapping quality (20), reads marked as duplicates, and reads with a bad CIGAR string (for further details, refer to the GATK documentation available at https://software.broadinstitute.org/gatk). Variants identified by SAMtools mpileup were required to have a minimum read mapping quality of 20 (MAPQ20) and a minimum base quality of 20. For each dataset, the genotypes of SNPs that were identified by both variant detection algorithms were retrieved from SAMtools, resulting in a total of 1,065,798 SNPs in Arabian Peninsula samples and 402,077 SNPs in the Sudanese populations. These variants were further hard-filtered using the VariantFiltration tool of GATK. This included removing variants with low quality by depth (QD < 2) to normalize variant quality and avoid inflation in the presence of deep coverage, removing variants indicating a high probability of strand bias (FS > 60), removing variants with a low root mean score mapping quality (MQ < 40), removing variants with low variant site quality (QUAL < 30), retaining variants where the mapping qualities of reads supporting the reference and alternate allele did not exhibit a bias for either allele (MQRankSum < −12.5), and retaining variants where their positions did not exhibit a bias toward the ends of reads (ReadPosRankSum < −8). The remaining variants were subsequently filtered using bcftools version 1.6 (Li, 2011) to retain those with a depth of coverage (DP) ≥ 5, and with a per-SNP DP within three standard deviations (SD) of the mean DP across all samples for a given dataset. The transition/transversion (Ts/Tv) ratio was calculated before and after filtering using bcftools stats. Variants on the mtDNA (scaffold NC_009849.1), indels, and SNPs that were not bi-allelic were also removed. A total of 206,415 and 1,028,936 SNPs were retained for the GBS and WGS data, respectively (**Supplementary Table S1**).

The filtered genotype data from the WGS and GBS datasets were merged, retaining a total of 1,173,266 SNPs common to both, and is herein referred to as the merged dataset. The filtered genotype data from the GBS dataset, independent of the WGS dataset, is herein referred to as the Sudanese dataset. A quality control (QC) steps was performed using the *check.marker* function implemented in the GenABEL package (Aulchenko et al., 2007) for R software version 2.15.1 (R Development Core Team, 2012). SNPs with minor allele frequency (MAF) less than 5% and call rate less than 95% were excluded from each dataset. A breakdown of the SNPs failing QC in each dataset is provided in **Table 2**. The final numbers of SNPs after QC were 12,920 and 39,843 in the merged and Sudanese datasets, respectively. For the genetic diversity analyses, SNPs with high linkage disequilibrium

TABLE 2 | Summary of SNPs numbers following quality control (QC) process.

	Datasets	
	Merged dataset	**Sudanese dataset**
	Raw SNP number	
	1,173,266	206,415
Quality control criteria		
MAF[1] < 5%	105,394	39,937
Call rate < 95%	1,158,312	156,472
Both MAF and call rate	103,360	29,837
Linkage disequilibrium (LD)[2]		26,804
	Final SNP number	
Before LD pruning	12,920	39,843
After LD pruning		13,039

[1] *Minor allele frequency.*
[2] *For genetic diversity analyses.*

(LD) ($r^2 > 0.1$) were filtered out using the *indep-pairwise* tool (–indep-pairwise 50 10 0.1) in PLINK version 1.9 (Purcell et al., 2007). A total of 7,273 and 26,804 SNPs were removed from the merged and Sudanese datasets, respectively. For the pooled heterozygosity (*Hp*) analyses of the Sudanese dataset, the MAF criteria were not applied as the statistic is specifically testing for large deviations in heterozygosity levels and as such is sensitive to MAF. The LD filter was also excluded, as regions under selection are expected to accommodate SNPs in LD, and so removing these reduces the power of the analysis. The QC filtering was applied to the combined Sudanese, packing, and racing camel subsets independently, resulting in 49,943 SNPs in the combined Sudanese, 48,808 SNPs in packing and 46,535 SNPs in racing camels. None of the samples from either QC-filtered dataset (merged and Sudanese) exhibited a genotyping call rate <95%, or an identity by state (IBS) ≥ 90% with any other sample.

Genetic Diversity Analyses

The mean heterozygosity and the inbreeding coefficient (*Fis*) of the different camel populations were computed from the merged dataset using the *hom* function implemented in GenABEL (Aulchenko et al., 2007). The two-sample Mann–Whitney U test was used to test for statistically significant differences in the heterozygosity and *Fis* values between the Arabian Peninsula and Sudanese camels, and between the different Sudanese populations. The one-sample Mann–Whitney U test was used to check if the *Fis* values of each of the distinct Sudanese populations, the combined Sudanese populations, and the Arabian Peninsula camels were significantly different from zero. Principle component analyses (PCA) were conducted in R using the *prcomp* function on both datasets to determine the genomic relationship between Arabian Peninsula and Sudanese camels, and separately among the Sudanese camels only.

To assess levels of genetic ADMIXTURE, analyses were performed using ADMIXTURE 1.23 (Alexander et al., 2009) on each dataset, assuming a number of clusters ranging from 1 to the number of populations sampled in each dataset (merged = 8; Sudanese = 5). In each case, 200 bootstrap iterations were performed. The optimal number of clusters was determined following Evanno et al. (2005) by calculating the second order rate of change of the likelihood for each *K* value. Figures of the ancestry assignments were plotted using the *ggplot2* package (Wickham, 2009) for R.

The relatedness of samples was evaluated in PLINK using the –make-rel tool. This tool estimates genetic relatedness among genome-wide SNPs based on the unadjusted A_{jk} statistic described in Yang et al. (2011), which ranges from 0 for an unrelated pair of individuals to 1 when comparing an individual to itself. Yang et al. (2011) consider a pair of individuals to be related where $A_{jk} > 0.025$. The relationship matrix was plotted using the ggplot2 package for R.

Signatures of Selection Analyses

Pooled Heterozygosity (Hp) Analysis

Hp values were calculated using the formula described by Rubin et al. (2010): $Hp = 2 \sum n_{MAJ} \sum n_{MIN} / (\sum n_{MAJ} + \sum n_{MIN})^2$. The analysis was performed in 50-kb sliding windows with a 25-kb step on all of the Sudanese camels, and independently for the Sudanese packing and racing types. After discarding windows supported by only a single SNP, the major and minor allele counts for each SNP were recorded (n_{MAJ} and n_{MIN}, respectively), and *Hp* values of each 50-kb window were calculated, which were subsequently Z-transformed: ZHp = *Hp* – median *Hp*/SD *Hp*. Values in the extreme lower 0.1% tail of the empirical distribution of each analysis were considered significant (all Sudanese Zhp = −2.51; packing Zhp = −2.45; racing Zhp = −2.22) (**Supplementary Figure S2**). Windows with significant Zhp values that overlapped were merged into single regions.

Fixation Index (*Fst*) Analysis

Fixation index (*Fst*) analysis (Weir and Cockerham, 1984) was calculated between Sudanese packing and racing camels in 50-kb windows with a 25-kb step using VCFtools version 0.1.13 (Danecek et al., 2011). As with the *Hp* analyses, windows comprising only a single SNP were discarded. Values in the extreme upper 0.1% tail of the *Fst* distribution were considered significant (*Fst* = 0.206) (**Supplementary Figure S2**). Windows with significant *Fst* values that overlapped were merged into single regions.

Functional Characterization of Candidate Regions of Selection

The positions of the significant windows from each analyses were cross-referenced against the genes annotated in the Arabian Peninsula dromedary camel reference genome assembly using the *intersectBed* function from the *BedTools* software (Quinlan and Hall, 2010). From each analysis, a background gene set was also identified by retrieving all genes that intersect with all windows tested. Gene-set over-representation analyses were performed using the ConsensusPathDB-human (CPDB) online tool (http://cpdb.molgen.mpg.de/CPDB). From each analysis, the significant and background gene lists were analyzed, and the significantly enriched biological pathways and gene ontology (GO) terms were recorded. Pathways and GO terms that remained significant after false discovery rate (FDR) correction (*q*-value < 0.05) were considered for discussion. A review of the literature for the genes identified from each analysis was also conducted to evaluate their relevance to adaptation to the desert environment, production, and performance traits.

RESULTS

Genetic Diversity and Relationship Among the Camel Populations

The mean observed heterozygosity in Arabian Peninsula (0.560 ± 0.003) was significantly higher than within Sudanese camel populations (0.347 ± 0.003) (*P*-value < 0.0001). Among Sudanese camels, heterozygosity values were not significantly different (*P*-values > 0.05) ranging from 0.341 in Anafi to 0.353 in Bishari (**Table 3**). The mean *Fis* value of Arabian Peninsula samples (−0.814 ± 0.01) was significantly lower than in Sudanese

TABLE 3 | Summary of observed heterozygosity and inbreeding coefficient (*Fis*) for each camel population.

Population	Mean observed heterozygosity	Heterozygosity standard deviation	Mean Fis	Fis standard deviation
Majaheem + Sofor (Alwatania camel dairy farm)	0.56	0.003	−0.811	0.01
Majaheem (the conservation and genetic improvement center)	0.562	0.002	−0.820	0.007
Omani	0.559	0.005	−0.810	0.016
Arabi	0.348	0.009	−0.126	0.029
Kenani	0.345	0.008	−0.118	0.026
Lahawi	0.348	0.009	−0.127	0.029
Anafi	0.341	0.010	−0.105	0.032
Bishari	0.353	0.015	−0.141	0.050
All Arabian Peninsula	0.560	0.003	−0.814	0.011
All Sudanese	0.347	0.010	−0.123	0.033

camel populations (−0.123 ± 0.033) (*P*-value < 0.0001). As with the heterozygosity values, the *Fis* values among the Sudanese camel populations were not significantly different from each other (*P*-value > 0.05) (**Table 3**). *Fis* values of the Arabian Peninsula and Sudanese dromedary populations were all significantly different from zero (*P*-value < 0.05) (**Supplementary Table S2**). The camels sampled from Arabian Peninsula exhibited higher relatedness (median $A_{jk} = 0.347 \pm 0.093$) than those sampled from Sudan (median $A_{jk} = -0.007 \pm 0.018$) (**Supplementary Figure S3**, **Supplementary Figure S4** and **Supplementary Table S3**).

Principal component analysis (PCA) of the merged dataset differentiates the camel populations of Sudan from the Arabian Peninsula along the first principal component (PC1), which explains 9.2% of the total variation. PC2, which accounts for 3.2% of the total variation, distinguishes the two Majaheem from Al-Kharj from the other Arabian Peninsula dromedary (**Figure 1A**), while PC3, which explains 2.8% of the total variation, separates the Omani camels from the other Arabian Peninsula dromedary (**Supplementary Figure S5**). Analysis of the Sudanese camels alone reveals PC1 and PC2 to explain only 3.4 and 3.2% of the total variation, respectively. This indicates that the populations sampled are relatively homogeneous. Nonetheless, the variance of data along these components indicates two Bishari and three Anafi to diverge from the origin for PC1 and PC2, respectively (**Figure 1B**).

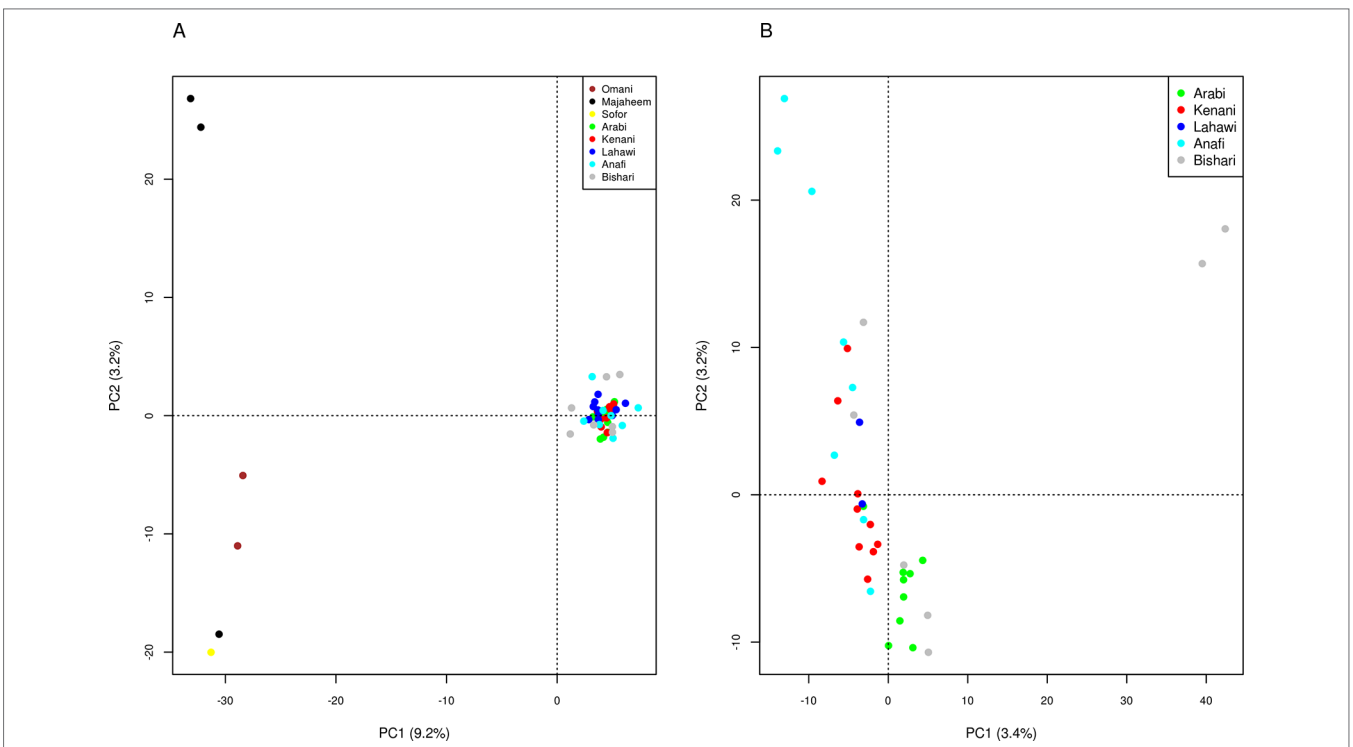


FIGURE 1 | Principle components analysis, components 1 and 2 for **(A)** camel populations from Sudan and the Arabian Peninsula and **(B)** camel populations from Sudan only.

We applied the ΔK approach of Evanno et al. (2005) to identify the optimal number of genetic backgrounds from the results of the ADMIXTURE analyses of the merged dataset and the Sudanese dataset. In both cases, the optimal number of genetic backgrounds is found to be two (**Supplementary Figure S6**). This is not surprising in the merged dataset, given the differentiation of samples collected from Sudan and the Arabian Peninsula by PCA (**Figure 1A**). At $K = 3$, a degree of ADMIXTURE is observed among the Sudanese camels, with two samples from Bishari providing the signal for the 3rd genetic background (**Figure 2**). This pattern of ADMIXTURE is broadly observed at $K = 2$ in the independent analysis of Sudanese camels (Sudanese dataset; **Supplementary Figure S7**), although the extent of ADMIXTURE is slightly larger. At $K < 4$,

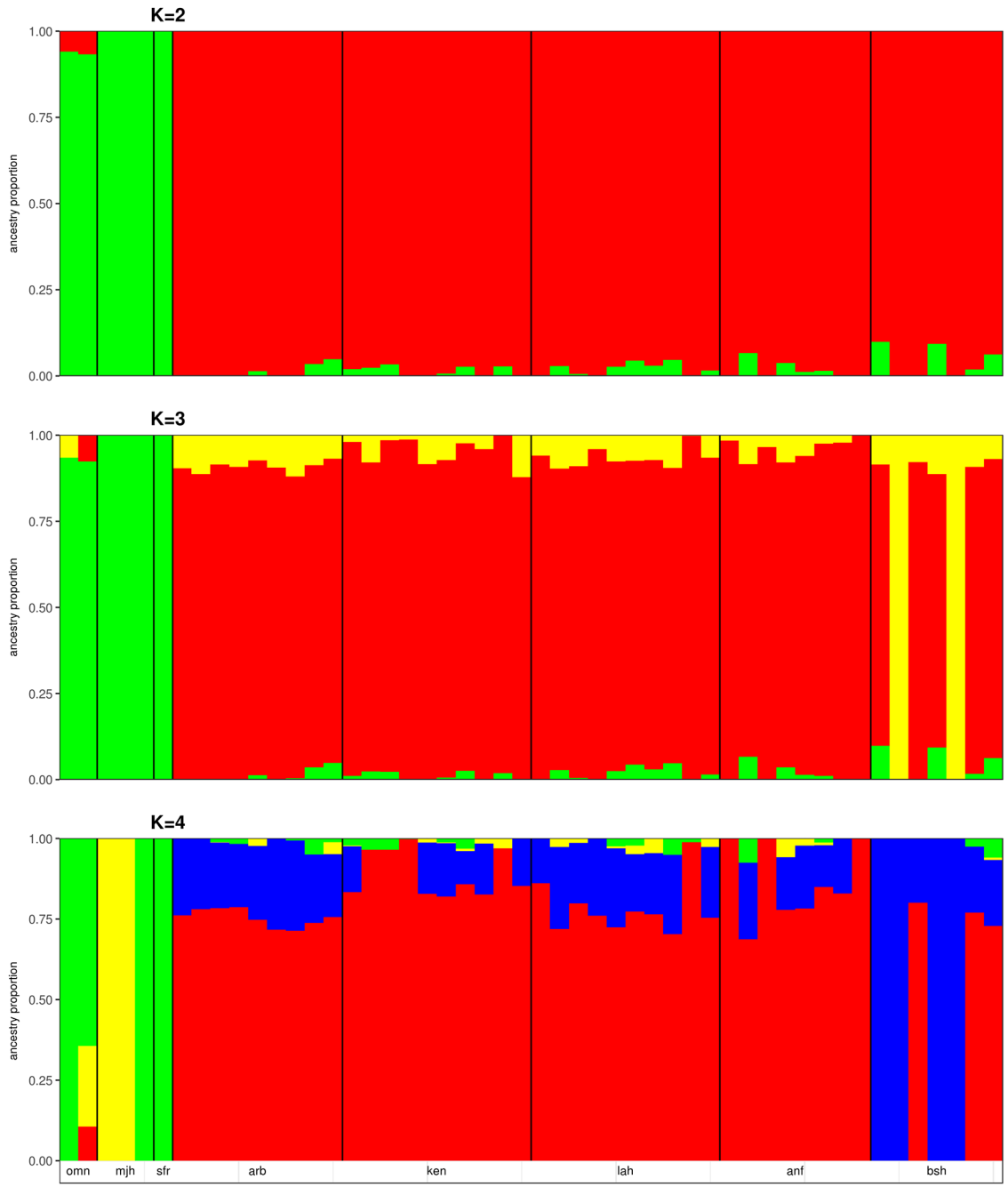


FIGURE 2 | ADMIXTURE plots of Sudanese and Arabian Peninsula camels for cluster values (*K*) from 2 to 4. omn, Omani; mjh, Majaheem; sfr, Sofor; arb, Arabi; ken, Kenani; lah, Lahawi; anf, Anafi; bsh, Bishari.

the camels sampled from the Arabian Peninsula exhibit a more homogenous genetic background, distinct from the Sudanese camels with traces of Sudanese ancestry found in Omani camels (**Figure 2**). This distinction can also be observed at *K* values from 5 to 7 (**Supplementary Figure S8**). At *K* = 4 and 8, the two Majaheem samples from the Conservation and Genetic Improvement Centre provide a separate genetic background, while a single Majaheem camel retains a background shared with the Sofor sample which is also the dominant background in the Omani camels (**Figure 2** and **Supplementary Figure S8**).

Candidate Signatures of Positive Selection

The mean *Hp* value across all sliding windows analyzed for the three Sudanese camel subsets (all Sudanese, packing, racing) was 0.24 ± 0.09. The analysis of all Sudanese camels returned 267 significant windows representing 176 regions. The subsets of packing and racing camels returned 281 and 472 significant windows representing 189 and 308 regions, respectively (**Table 4** and **Supplementary Table S4**), of which 133 were common to both camel types (**Supplementary Table S5**). A total of 159 and 148 regions identified in packing and racing camels, respectively, overlapped with the significant regions identified across all Sudanese camels. In total, 132 regions were found to overlap across each analysis (**Supplementary Table S5**).

The *Fst* analysis revealed 13 out of 13,097 windows to exhibit high genetic differentiation (*Fst* > 0.206) between racing and packing camels from Sudan. These windows represented eight distinct regions (**Table 4** and **Supplementary Table S6**).

Functional Characterization of Candidate Regions of Selection

The *Hp* analysis identified 132, 137, and 255 genes to intersect with 86, 99, and 161 regions in all Sudanese, packing, and racing camels, respectively (**Table 4** and **Supplementary Table S7**). While the *Fst* analysis identified 11 genes to overlap with five genetically differentiated regions between Sudanese packing and racing camels (**Table 4** and **Supplementary Table S8**). These genes are associated with different biological pathways, including immune response, fertility, milk content, energy homeostasis, and chondrogenesis (**Table 5**). No biological pathways were found to be significantly over-represented in the genes identified from the analyses of all Sudanese and packing camels. While a single GO term (monocarboxylic acid transmembrane transporter activity) was found to be significantly enriched in the analysis of all Sudanese camels. Moreover, a single biological pathway associated with liver kinase B1 (*LKB-1*) signaling was found to be significantly over-represented in Sudanese racing camels (**Table 6** and **Supplementary Table S9**). In total, four biological pathways and five GO terms are significantly enriched among the genes identified from the *Fst* analysis between Sudanese racing and packing camels (**Table 6** and **Supplementary Table S9**).

TABLE 4 | Summary of windows, regions, and genes identified in signatures of selection analyses in Sudanese camel.

	Signatures of selection analysis			
	Hp			*Fst*
	All Sudanese camels	Packing camels	Racing camels	
Number of windows	267	281	472	13
Number of merged regions	176	189	308	8
Number of genes	132	137	255	11
Number of regions with genes	86	99	161	5

TABLE 5 | Genes intersecting the pooled heterozygosity (*Hp*) and genetic differentiation (*Fst*) candidate regions.

Functional category	Gene ID	Gene description	Type of candidate region
Immune response	C9	Complement component C9	Packing/racing/all Sudan camels (Hp region)
	IL6R	Interleukin-6 receptor subunit alpha	Packing/racing/all Sudan camels (Hp region)
	CCR8	C–C chemokine receptor type 8	Packing/all Sudan camels (Hp region)
	CX3CR1	CX3C chemokine receptor 1	Packing/racing/all Sudan camels (Hp region)
	LOC105100014	Complement receptor type 1-like	Racing camels/all Sudan camels (Hp region)
	C1QTNF8	Complement C1q tumor necrosis factor-related protein 8	Packing/all Sudan camels (Hp region)
Fertility	LOC105094930	Olfactory receptor 1S1-like	Racing camels (Hp region)
	LOC105094932	Olfactory receptor 5B12	Racing camels (Hp region)
	LOC105094933	Olfactory receptor 5B3-like	Racing camels (Hp region)
	ESR1	Estrogen receptor	Racing camels (Hp region)
	SPACA5	Sperm acrosome-associated protein 5	Packing/racing/all Sudan camels (Hp region)
Milk content	PICALM	Phosphatidylinositol-binding clathrin assembly protein	Packing/racing/all Sudan camels (Hp region)
Chondrogenesis	LOC105087163	Chondroitin sulfate proteoglycan 4-like	Packing/racing/all Sudan camels (Hp region)
	CRLF1	Cytokine receptor-like factor 1	Fst region
	CHSY1	Chondroitin sulfate synthase 1	Racing (Hp region)
Energy homeostasis	ESRRG	Estrogen-related receptor gamma	Packing/racing/all Sudan camels (Hp region)
	CRTC1	CREB-regulated transcription coactivator 1	Fst region
Running performance	NAA16	N (alpha)-acetyl transferase 16	Fst region

DISCUSSION

In this study, we investigated the genetic diversity and relationship between camels sampled from Sudan and the Arabian Peninsula using genotype data derived from GBS and WGS. The migration of camels to Sudan from their putative center of domestication in the Arabian Peninsula may have resulted in a founder effect coupled with genetic drift that can explain the reduced heterozygosity observed in the Sudanese camels we studied. Populations that are closer to their centers of domestication are expected to exhibit higher levels of heterozygosity than more distant populations. Thus, the higher level of heterozygosity observed in the camels sampled from the Arabian Peninsula compared to those from Sudan offers some support for the southeast of the Arabian Peninsula as a center of domestication. Investigating a greater breadth and depth of sampling throughout Africa and the Arabian Peninsula will provide further context for these interpretations.

The Arabian Peninsula and Sudanese camel populations both exhibit mean negative *Fis* values significantly different from zero, which may be suggestive of low levels of inbreeding among these populations as a consequence of their historical use in transportation and trading—which might be associated with continuous interbreeding and gene flow. We sought to evaluate the relatedness of samples using the A_{jk} statistic, which indicated the majority of Sudanese camels to be unrelated, whereas all of the camels from the Arabian Peninsula returned high values ($A_{jk} > 0.025$). These high values observed across the Arabian Peninsula camels are likely an artifact of the small sample size that can result in severely biased estimates (Wang, 2017).

PCA and ADMIXTURE analyses revealed a phylogeographic distinction between the Arabian Peninsula and Sudanese camels. This might be attributed to the geographical isolation of these two populations by the Red Sea, hindering gene flow between them. Such genetic distinction has been observed between camel populations from Kenya, the Arabian Peninsula, and Pakistan (Mburu et al., 2003). We also observed traces of Sudanese ancestry in Omani camels, which might possibly result from historical interbreeding while following the trading routes connecting Africa with the south of Arabian Peninsula (Andree-Salvini, 2010). These results imply that the history of dromedary camel migration and interbreeding is highly complex and requires further investigation. A recent study by Almathen et al. (2016) showed that Sudanese camels are genetically distinct to camels from the south of the Arabian Peninsula but not the north. The discrepancy between that study and our results is likely to be due to our limited sampling of the Arabian Peninsula, which did not adequately capture the broader diversity of camel populations present. Furthermore, Almathen et al. (2016) employed autosomal microsatellite markers, whereas our genotypic data was derived from next-generation sequencing and is potentially more informative.

We observed no clear genetic structuring of the different Sudanese camel populations studied, providing further support of continuous gene flow. The two Bishari and three Anafi divergent camels exhibited marginally higher relatedness with one another than with other camels sampled from the same regions (**Supplementary Figure S3**). This likely accounts for their divergence in the PCA (**Figure 1B**). The lack of genetic distinction between the packing and the racing camels reflects a common practice in Sudan of breeding racing camels from East Sudan with packing camels in the West to improve the performance of packing camels. This is also reflected in the mean negative *Fis* values in the Sudanese dromedary camel populations analyzed.

The camels from the Arabian Peninsula were broadly found to be homogeneous in the ADMIXTURE analyses. PCA differentiated Omani samples from the rest of the Arabian Peninsula along the 3rd principal component. The 2nd principle component, although explaining little variation, differentiated the Majaheem and Sofor samples from the Alwatania camel dairy farm from the two Majaheem samples from the conservation and genetic improvement center. This was also observed in the

TABLE 6 | Summary of significantly enriched biological pathways and gene ontology (GO) terms following gene-set over-representation analyses.

	Signatures of selection analysis						
	Hp					*Fst*	
	All Sudanese camels	***q*-value[1]**	**Packing camels**	**Racing camels**	***q*-value[1]**		***q*-value[1]**
Biological pathway	None	None	None	LKB1 signaling events	0.015	Brain-derived neurotrophic factor (BDNF) signaling pathway	0.006
						Human T-cell leukemia virus 1 infection—Homo sapiens (human)	0.006
						Signaling by interleukins	0.006
						Cytokine signaling in immune system	0.011
Gene ontology (GO) term	Monocarboxylic acid transmembrane transporter activity	0.019	None	None		Ribosome binding	0.001
						Ureteric bud development	0.032
						Mesonephros development	0.032
						Kidney epithelium development	0.032
						Ribonucleoprotein complex binding	0.007

[1] *The q-values in all analyses are derived from false discovery rate correction.*

ADMIXTURE analysis at K = 4 and K = 8. This suggests that camels from the Arabian Peninsula might exhibit a degree of genetic structure based on their breeding location and/or geographical origin. This observation however, which agrees with the findings of Almathen et al. (2016), requires further investigation using a larger and more diverse sample of camels from the Arabian Peninsula.

We further investigated the Sudanese camels for signatures of positive selection. A number of genomic regions exhibiting reduced heterozygosity were identified across the Sudanese camels, and within the subsets of packing and racing camels (**Table 4**). The *Fst* analysis, however, revealed few genetically differentiated regions between racing and packing camels. There was no overlap between the regions identified by the *Fst* analysis and those identified by the *Hp* analysis. This is not entirely unexpected as the two analyses are better-suited to different selection time frames, whereby *Fst* focuses on more recent selection than the *Hp* statistic (Oleksyk et al., 2010).

The regions identified from these analyses host a number of genes, for which gene set over-representation analyses and extensive literature review revealed a number to be of functional interest. In particular, we identified genes linked to the immune response, fertility, milk content, chondrogenesis, energy homeostasis, and running performance (**Table 5**). Whether or not these genes are linked to the domestication of dromedary camels from their wild ancestors, as found in sheep and goats (Alberto et al., 2018), requires further validation using genomic data from the wild ancestor of dromedary camels.

Examples of genes involved in the immune response include members of the complement system (*C9, LOC105100014*, and *C1QTNF8*), CX3C chemokine receptor 1 (*CX3CR1*), and interleukin-6 receptor subunit alpha (*IL6R*) genes, both of which were identified in packing and racing camels. The complement system links the innate and adaptive immune responses, mediating responses to inflammatory triggers through a co-ordinated enzyme cascade (Nesargikar et al., 2012). *CX3CR1* and interleukin-6 (*IL-6*) are both involved in inducing inflammation in response to infection and tissue injuries (Ishida et al., 2008; Tanaka et al., 2014), while *IL-6* is also involved in native T-cell differentiation (Tanaka et al., 2014). The chemokine receptor type 8 (*CCR8*) gene, identified in packing camels, plays a role in regulating the immune system by controlling regulatory T-cell activity (Coghill et al., 2013). Genes in this category have also been identified to be under adaptive evolution in dromedary and Bactrian camels in a study by Wu et al. (2014).

Genes involved in fertility have been identified in both packing and racing Sudanese dromedary camels. These include the *ESR1* gene which encodes the estrogen receptor alpha that mediates estrogen action on target tissues. Mice whose estrogen receptor-alpha gene has been knocked out demonstrate complete infertility (Matthews and Gustafsson, 2003) and impaired spermatogenesis (Couse et al., 2001). Olfactory receptors have been shown to be expressed in mature sperm (Parmentier et al., 1992; Vanderhaeghen et al., 1993) and to play a role in oocyte fertilization upon interaction with chemo-attractants secreted from oocyte-cumulus cells (Spehr et al., 2003; Fukuda et al., 2004). The sperm acrosome-associated protein 5 gene (*SPACA5*) plays a role in acrosome reaction and the fertilization process (Agarwal et al., 2015). Fertility-associated genes have also been found to be under selection in indigenous African zebu cattle (Bahbahani et al., 2018a; Bahbahani et al., 2018b) in order to maintain fertility in the harsh environment (Skinner and Louw, 1966; Hansen, 2004).

An example of a gene likely to be a feature of adaptation to the desert environment is *ESRRG*. This gene, which is in candidate regions identified in both camel types, is related to energy homeostasis. *ESRRG* encodes the estrogen-related receptor gamma protein, which is involved in regulating metabolism and energy production in cells (Eichner and Giguere, 2011). A transition polymorphism (rs1890552 A > G) in this gene has been found to be associated with decreased levels of fasting glucose in humans (Kim et al., 2017b). The gene-set over-representation analysis in racing camel also identified the over-representation of genes associated with *LKB1* signaling pathway, which is associated with regulating cellular energy metabolism in eukaryotic cells (Alessi et al., 2006). Genes in this category have also been found to be under adaptive evolution in dromedary and Bactrian camels (Wu et al., 2014).

A gene which encodes the phosphatidylinositol-binding clathrin assembly protein (*PICALM*), which is involved in regulating milk content, was identified in both racing and packing camel types. Variants in *PICALM* have been associated with α_{S1}-casein content in cattle milk (Sanchez et al., 2017). This gene is therefore of potential future interest to improve the productivity of dromedary camels through genomic selection breeding programs (Al Abri and Faye, 2019).

A number of genes identified in candidate regions in packing and racing camels were found to be associated with chondrogenesis. These include the *LOC105087163* gene which encodes the chondroitin sulfate proteoglycan 4 protein. Members of this protein family, such as aggrecan and versican, have been found to be involved in cartilage and limb joint formation (Watanabe et al., 1994; Choocheep et al., 2010). Another chondrogenesis-related gene identified, *CHSY1*, encodes the chondroitin sulfate synthase 1 protein, which, when knocked-out in mice, results in cartilage impairment, aberrant joint formation, and decreased bone density (Wilson et al., 2012). The identification of genes associated with chondrogenesis in packing dromedary camels might be a reflection of the historical use of camels in trading and transportation, which likely placed chronic physical demands on these animals similar to other livestock employed in the provision of draught power. The signatures of selection identified for genes associated with dairy traits in racing camels might also be a historical reflection of the general use of camels by nomads in the provision of milk.

A number of genes associated with energy homeostasis, chondrogenesis, and running performance were identified within the genetically differentiated regions between racing and packing camels. One such example is CREB-regulated transcription coactivator 1 (*CRTC1*). This gene is linked to energy homeostasis, which is likely to be an important function in racing camels in order to maintain their stamina during racing competitions. Studies in mice support this, indicating that the gene is associated with energy balance (Altarejos et al., 2008). Another candidate gene identified was cytokine receptor-like factor 1 (*CRLF1*), which plays a role in chondrogenesis. The

expression of this gene has been shown to be up-regulated in mice chondrocytes upon stimulation by TGF-beta factor which, together with its co-factor cardiotrophin-like cytokine factor 1 (CLCF1), induces proliferation of chondrocyte precursors (Stefanovic and Stefanovic, 2012). A final example, associated with running performance, is the N(alpha)-acetyl transferase 16 (*NAA16*) gene, which has been correlated with average running speed in mice (Kelly et al., 2014). Interestingly, three out of the five significantly enriched gene ontology terms in the *Fst* analysis genes are associated urinary tract and kidney development, which might be linked with the running performance of racing camels (**Table 6**). A study by Leikis et al. (2006) has revealed reduced exercise performance in patients with chronic kidney disease at stages 3 and 4. This reduction in exercise capacity, and associated muscle strength, was also been linked with renal function failure. Moreover, the impairment in muscle K^+ level regulation, which is associated with renal function failure (Bergstrom et al., 1983), has also been found to be linked with muscle fatigue (Sejersted and Sjogaard, 2000). This requires further investigations through detailed physiological studies on racing and packing dromedary camels.

To further our understanding of the genetic mechanisms underlying phenotypic traits of importance in dromedary camel, a number of actions are required. In the first instance, the reference genome is currently at the scaffold stage and would greatly benefit from the application of long-read single-molecule sequencing technologies coupled with optical mapping to improve its quality (Mostovoy et al., 2016; Weissensteiner et al., 2017). Improving the assembly of the reference genome to reach the chromosome stage is essential to facilitate detailed analyses for signatures of selection. The recent construction of two whole-genome radiation hybrid panels for dromedary camels by Perelman et al. (2018) is a promising first step toward improving the assembly. Secondly, in parallel to work improving the contiguity of the reference genome is the need for functional analyses on the annotated genes. Improving our understanding of the function of genes provides greater context toward interpreting genotype–phenotype associations. Thirdly, the detailed characterization of camels using standard phenotypic and morphometric parameters is required to more accurately classify the different camel types. Using WGS data in future studies instead of GBS will provide greater resolution to analyses, and less redundancy of data, while increasing the breadth and depth of sampling will enable more reliable calculations of genetic diversity parameters.

CONCLUSION

We have reported here for the first time the phylogeographic classification between the Arabian Peninsula and Sudanese dromedary camel populations using WGS and GBS data. We identified a number of genomic regions under positive selection hosting genes of putative relevance to Sudanese packing and racing camels, including genes potentially associated with dairy traits and running performance. The results of this study call for further investigation of the genome of dromedary camel using larger and more diverse populations to better identify variants associated with their important phenotypes. The results of which can support the development of informed breeding programs with the aim to improve the productivity and performance of dromedary camels.

DATA AVAILABILITY

The Arabian Peninsula camel full genome sequence data and the genotyping-by-sequencing data of Sudanese camels are publicly available from the European Nucleotide Archive (ENA) with the Bioproject accession number PRJEB32117.

ETHICS STATEMENT

Standard techniques were used to collect blood. The procedure was reviewed and approved by Faculty of Veterinary Science, University of Nyala, Sudan.

AUTHOR CONTRIBUTIONS

HB, HM and OH conceived and designed the project. ES collected the dromedary blood samples from Sudan. HM and FA contributed sequence data for the Sudanese and Arabian Peninsula samples, respectively. HB and DW performed bioinformatic analysis. HB, DW and OH wrote the manuscript. All authors have commented upon and agreed on the contents of the manuscript.

FUNDING

This work is part of the project "Agricultural growth, capacity building for scientific preservation of livestock breeds in Sudan." The project was supported by a grant from the Korea-Africa Economic Cooperation Trust Fund through the African Development Bank and co-funded by the livestock CGIAR-CRP. We would like to extend our gratitude to King Faisal University (Saudi Arabia) for their research award (Grant F289) to sequence the genomes of the six Arabian Peninsula dromedary camel samples.

ACKNOWLEDGMENTS

We would like to extend our sincere gratitude to the undersecretary of the Ministry of Finance and National Economy, Sudan, and the resident representative(s) of the African Development Bank (Khartoum, Sudan) for their support in the approval and implementation of the project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00893/full#supplementary-material

REFERENCES

Al Abri, M. A., and Faye, B. (2019). Genetic improvement in dromedary camels: challenges and opportunities. *Front. Genet.* 10, 167. doi: 10.3389/fgene.2019.00167

Agarwal, A., Sharma, R., Durairajanayagam, D., Ayaz, A., Cui, Z., Willard, B., et al. (2015). Major protein alterations in spermatozoa from infertile men with unilateral varicocele. *Reprod. Biol. Endocrinol.: RB&E* 13, 8–8. doi: 10.1186/s12958-015-0007-2

Al-Swailem, A. M., Al-Busadah, K. A., Shehata, M. M., Al-Anazi, I. O., and Askari, E. (2007). Classification of Saudi Arabian camel (Camelus dromedarius) subtypes based on RAPD technique. *J. Food Agric. Environ.* 5, 143. doi: 10.1234/4.2007.749

Alberto, F. J., Boyer, F., Orozco-Terwengel, P., Streeter, I., Servin, B., De Villemereuil, P., et al. (2018). Convergent genomic signatures of domestication in sheep and goats. *Nat. Commun.* 9, 813. doi: 10.1038/s41467-018-03206-y

Alessi, D. R., Sakamoto, K., and Bayascas, J. R. (2006). LKB1-dependent signaling pathways. *Annu. Rev. Biochem.* 75, 137–163. doi: 10.1146/annurev.biochem.75.103004.142702

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Ali, A., Baby, B., and Vijayan, R. (2019). From desert to medicine: a review of camel genomics and therapeutic products. *Front. Genet.* 10, 17. doi: 10.3389/fgene.2019.00017

Almathen, F., Charruau, P., Mohandesan, E., Mwacharo, J. M., Orozco-Terwengel, P., Pitt, D., et al. (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc. Natl. Acad. Sci. U. S. A.* 113, 6707–6712. doi: 10.1073/pnas.1519508113

Almathen, F., Elbir, H., Bahbahani, H., Mwacharo, J., and Hanotte, O. (2018). Polymorphisms in MC1R and ASIP genes are associated with coat colour variation in the Arabian camel. *J. Hered.* 109, 700–706. doi: 10.1093/jhered/esy024

Andree-Salvini, B. (2010). *Roads of Arabia: archeology and history of the Kingdom of Saudi Arabia*. Paris, France: Somogy Editions d'Art.

Altarejos, J. Y., Goebel, N., Conkright, M. D., Inoue, H., Xie, J., Arias, C. M., et al. (2008). The Creb1 coactivator Crtc1 is required for energy balance and fertility. *Nat. Med.* 14, 1112–1117. doi: 10.1038/nm.1866

Aulchenko, Y. S., Ripke, S., Isaacs, A., and Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294–1296. doi: 10.1093/bioinformatics/btm108

Babar, M. E., Hussain, T., Wajid, A., Nawaz, A., Nadeem, A., Shah, S. A., et al. (2015). Mitochondrial cytochrome-b and D-loop sequence based genetic diversity in Mareecha and Bareela camel breeds of Pakistan. *J. Anim. Plant Sci.* 25, 591–594. doi: 10.1111/age.12158

Bahbahani, H., Afana, A., and Wragg, D. (2018a). Genomic signatures of adaptive introgression and environmental adaptation in the Sheko cattle of southwest Ethiopia. *PLoS One* 13, e0202479. doi: 10.1371/journal.pone.0202479

Bahbahani, H., Clifford, H., Wragg, D., Mbole-Kariuki, M. N., Van Tassell, C., Sonstegard, T., et al. (2015). Signatures of positive selection in East African Shorthorn Zebu: a genome-wide single nucleotide polymorphism analysis. *Sci. Rep.* 5, 11729. doi: 10.1038/srep11729

Bahbahani, H., Salim, B., Almathen, F., Al Enezi, F., Mwacharo, J. M., and Hanotte, O. (2018b). Signatures of positive selection in African Butana and Kenana dairy zebu cattle. *PLoS One* 13, e0190446. doi: 10.1371/journal.pone.0190446

Bahbahani, H., Tiijani, A., Mukasa, C., Wragg, D., Almathen, F., Nash, O., et al. (2017). Signature of selection for environmental adaptation and zebu x taurine hybrid fitness in East African Shorthorn Zebu. *Front. Genet.* 8, 1–20. doi: 10.3389/fgene.2017.00068

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., et al. (2013). Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* 193, 1073–1081. doi: 10.1534/genetics.112.147710

Bergstrom, J., Alvestrand, A., Furst, P., Hultman, E., and Widstam-Attorps, U. (1983). Muscle intracellular electrolytes in patients with chronic uremia. *Kidney Int. Suppl.* 16, S153–S160.

Choocheep, K., Hatano, S., Takagi, H., Watanabe, H., Kimata, K., Kongtawelert, P., et al. (2010). Versican facilitates chondrocyte differentiation and regulates joint morphogenesis. *J. Biol. Chem.* 285, 21114–21125. doi: 10.1074/jbc.M109.096479

Coghill, J. M., Fowler, K. A., West, M. L., Fulton, L. M., Van Deventer, H., Mckinnon, K. P., et al. (2013). CC chemokine receptor 8 potentiates donor Treg survival and is critical for the prevention of murine graft-versus-host disease. *Blood* 122, 825–836. doi: 10.1182/blood-2012-06-435735

Couse, J. E., Mahato, D., Eddy, E. M., and Korach, K. S. (2001). Molecular mechanism of estrogen action in the male: insights from the estrogen receptor null mice. *Reprod. Fertil. Dev.* 13, 211–219. doi: 10.1071/RD00128

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Eichner, L. J., and Giguere, V. (2011). Estrogen related receptors (ERRs): a new dawn in transcriptional control of mitochondrial gene networks. *Mitochondrion* 11, 544–552. doi: 10.1016/j.mito.2011.03.121

Eisa, M. O., and Mustafa, A. (2011). Production systems and dairy production of Sudan Camel (Camelus dromedarius): a review. *Middle-East J. Sci. Res.* 7, 132–135. doi: 10.12895/jaeid.20132.166

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Epstein, H., and Mason, I. L. (1971). *The origin of the domestic animals of Africa*. New York: Africana publishing corporation.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

FAO. The Food and Agriculture Organization of the United Nations (2019). Italy: FAOSTAT.

Faye, B., Abdelhadi, O. M. A., Ahmed, A. I., and Bakheit, S. A. (2011). Camel in Sudan: future prospects. *Livestock Res. Rural Dev.* 23, 219.

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2015). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16, 314–324. doi: 10.1111/1755-0998.12443

Fukuda, N., Yomogida, K., Okabe, M., and Touhara, K. (2004). Functional characterization of a mouse testicular olfactory receptor and its role in chemosensing and in regulation of sperm motility. *J. Cell Sci.* 117, 5835–5845. doi: 10.1242/jcs.01507

Gauthier-Pilters, H. D., and Anne, I. (1983). *The camel: its evolution, ecology, behavior, and relationship to man*. Chicago: University of Chicago Press.

Hansen, P. J. (2004). Physiological and cellular adaptations of zebu cattle to thermal stress. *Anim. Reprod. Sci.* 82-83, 349–360. doi: 10.1016/j.anireprosci.2004.04.011

Hashim, M. W., Galal, Y. M., Ali, M. A., Khalafalla, A. I., A., Hamid, A. S., et al. (2015). Dromedary camels in Sudan, types and sub types, distribution and movement. *Int. J. Pharm. Res. Anal.* 5, 8–12.

He, J., Zhao, X., Laroche, A., Lu, Z. X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* 5, 484. doi: 10.3389/fpls.2014.00484

Ishida, Y., Gao, J. L., and Murphy, P. M. (2008). Chemokine receptor CX3CR1 mediates skin wound healing by promoting macrophage and fibroblast accumulation and function. *J. Immunol.* 180, 569–579. doi: 10.4049/jimmunol.180.1.569

Kelly, S. A., Nehrenberg, D. L., Hua, K., Garland, T., Jr., and Pomp, D. (2014). Quantitative genomics of voluntary exercise in mice: transcriptional analysis and mapping of expression QTL in muscle. *Physiol. Genomics* 46, 593–601. doi: 10.1152/physiolgenomics.00023.2014

Khalkhali-Evrigh, R., Hafezian, S. H., Hedayat-Evrigh, N., Farhadi, A., and Bakhtiarizadeh, M. R. (2018). Genetic variants analysis of three dromedary camels using whole genome sequencing data. *PLoS One* 13, e0204028. doi: 10.1371/journal.pone.0204028

Kijas, J. W., Porto-Neto, L., Dominik, S., Reverter, A., Bunch, R., Mcculloch, R., et al. (2014). Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Anim. Genet.* 45, 754–757. doi: 10.1111/age.12197

Kim, E. S., Elbeltagy, A. R., Aboul-Naga, A. M., Rischkowsky, B., Sayre, B., Mwacharo, J. M., et al. (2016). Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. *Heredity (Edinb.)* 116, 255–264. doi: 10.1038/hdy.2015.94

Kim, J., Hanotte, O., Mwai, O. A., Dessie, T., Bashir, S., and Diallo, B. (2017a). The genome landscape of indigenous African cattle. *Genome Biol.* 18, 34. doi: 10.1186/s13059-017-1153-y

Kim, M., Yoo, H. J., Kim, M., Seo, H., Chae, J. S., Lee, S. H., et al. (2017b). Influence of estrogen-related receptor gamma (ESRRG) rs1890552 A > G polymorphism on changes in fasting glucose and arterial stiffness. *Sci. Rep.* 7, 9787. doi: 10.1038/s41598-017-10192-6

Leikis, M. J., McKenna, M. J., Petersen, A. C., Kent, A. B., Murphy, K. T., Leppik, J. A., et al. (2006). Exercise performance falls over time in patients with chronic kidney disease despite maintenance of hemoglobin concentration. *Clin. J. Am. Soc. Nephrol.* 1, 488–495. doi: 10.2215/CJN.01501005

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Magee, P. (2015). When was the dromedary domesticated in the ancient Near East? *Z. Orient-Archäologie* 8, 253–278.

Matthews, J., and Gustafsson, J. A. (2003). Estrogen signaling: a subtle balance between ER alpha and ER beta. *Mol. Interv.* 3, 281–292. doi: 10.1124/mi.3.5.281

Mburu, D. N., Ochieng, J. W., Kuria, S. G., Jianlin, H., Kaufmann, B., Rege, J. E., et al. (2003). Genetic diversity and relationships of indigenous Kenyan camel (Camelus dromedarius) populations: implications for their classification. *Anim. Genet.* 34, 26–32. doi: 10.1046/j.1365-2052.2003.00937.x

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., et al. (2016). A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat. Methods* 13, 587. doi: 10.1038/nmeth.3865

Musa, H. H., Shuiep, E. S., and El-Zubeir, I. (2006). Camel husbandry among pastoralists in Darfur, Western Sudan. *Nomadic Peoples* 10, 101–105. doi: 10.3167/082279406780246438

Nesargikar, P. N., Spiller, B., and Chavez, R. (2012). The complement system: history, pathways, cascade and inhibitors. *Eur. J. Microbiol. Immunol. (Bp)* 2, 103–111. doi: 10.1556/EuJMI.2.2012.2.2

Oleksyk, T. K., Smith, M. W., and O'brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 365, 185–205. doi: 10.1098/rstb.2009.0219

Onzima, R. B., Upadhyay, M. R., Doekes, H. P., Brito, L. F., Bosse, M., Kanis, E., et al. (2018). Genome-wide characterization of selection signatures and runs of homozygosity in Ugandan Goat Breeds. *Front. Genet.* 9, 318. doi: 10.3389/fgene.2018.00318

Parmentier, M., Libert, F., Schurmans, S., Schiffmann, S., Lefort, A., Eggerickx, D., et al. (1992). Expression of members of the putative olfactory receptor gene family in mammalian germ cells. *Nature* 355, 453–455. doi: 10.1038/355453a0

Pauciullo, A., Shuiep, E. S., Cosenza, G., Ramunno, L., and Erhardt, G. (2013). Molecular characterization and genetic variability at κ-casein gene (CSN3) in camels. *Gene* 513, 22–30. doi: 10.1016/j.gene.2012.10.083

Perelman, P. L., Pichler, R., Gaggl, A., Larkin, D. M., Raudsepp, T., Alshanbari, F., et al. (2018). Construction of two whole genome radiation hybrid panels for dromedary (Camelus dromedarius): 5000RAD and 15000RAD. *Sci. Rep.* 8, 1982. doi: 10.1038/s41598-018-20223-5

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5, 92–102. doi: 10.3835/plantgenome2012.05.0005

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

R Development Core Team (2012). *R: a language and environment for statistical computing*. Vienna, Austria.

Rincon, G., Weber, K. L., Eenennaam, A. L., Golden, B. L., and Medrano, J. F. (2011). Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *J. Dairy Sci.* 94, 6116–6121. doi: 10.3168/jds.2011-4764

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591. doi: 10.1038/nature08832

Sanchez, M.-P., Govignon-Gion, A., Croiseau, P., Fritz, S., Hozé, C., Miranda, G., et al. (2017). Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genet. Sel. Evol. GSE* 49, 68–68. doi: 10.1186/s12711-017-0344-z

Schmidt-Nielsen, K. (1959). The physiology of the camel. *Sci. Am.* 201, 140–151. doi: 10.1038/scientificamerican1259-140

Sejersted, O. M., and Sjogaard, G. (2000). Dynamics and consequences of potassium shifts in skeletal muscle and heart during exercise. *Physiol. Rev.* 80, 1411–1481. doi: 10.1152/physrev.2000.80.4.1411

Skinner, J. D., and Louw, G. N. (1966). Heat stress and spermatogenesis in Bos indicus and Bos taurus cattle. *J. Appl. Physiol.* 21, 1784–1790. doi: 10.1152/jappl.1966.21.6.1784

Shuiep, E. T. S., Giambra, I. J., El Zubeir, I. E. Y. M., and Erhardt, G. (2013). Biochemical and molecular characterization of polymorphisms of αs1-casein in Sudanese camel (Camelus dromedarius) milk. *Int. Dairy J.* 28, 88–93. doi: 10.1016/j.idairyj.2012.09.002

Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J. A., Wetzel, C. H., Zimmer, R. K., et al. (2003). Identification of a testicular odorant receptor mediating human sperm chemotaxis. *Science* 299, 2054–2058. doi: 10.1126/science.1080376

Stefanovic, L., and Stefanovic, B. (2012). Role of cytokine receptor-like factor 1 in hepatic stellate cells and fibrosis. *World J. Hepatol.* 4, 356–364. doi: 10.4254/wjh.v4.i12.356

Tanaka, T., Narazaki, M., and Kishimoto, T. (2014). IL-6 in inflammation, immunity, and disease. *Cold Spring Harbor Perspect. Biol.* 6, a016295. doi: 10.1101/cshperspect.a016295

Vanderhaeghen, P., Schurmans, S., Vassart, G., and Parmentier, M. (1993). Olfactory receptors are displayed on dog mature sperm cells. *J. Cell Biol.* 123, 1441–1452. doi: 10.1083/jcb.123.6.1441

Wang, K., Wu, P., Yang, Q., Chen, D., Zhou, J., Jiang, A., et al. (2018). Detection of selection signatures in Chinese Landrace and Yorkshire pigs based on genotyping-by-sequencing data. *Front. Genet.* 9, 119. doi: 10.3389/fgene.2018.00119

Wang, J. (2017). Estimating pairwise relatedness in a small sample of individuals. *Heredity (Edinb.)* 119, 302–313. doi: 10.1038/hdy.2017.52

Wardeh, M. F. (2004). Classification of the dromedary camels. *J. Camel Sci.* 1, 1–7.

Watanabe, H., Kimata, K., Line, S., Strong, D., Gao, L. Y., Kozak, C. A., et al. (1994). Mouse cartilage matrix deficiency (cmd) caused by a 7 bp deletion in the aggrecan gene. *Nat. Genet.* 7, 154–157. doi: 10.1038/ng0694-154

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x

Weissensteiner, M. H., Pang, A. W. C., Bunikis, I., Hoijer, I., Vinnere-Petterson, O., Suh, A., et al. (2017). Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res.* 27, 697–708. doi: 10.1101/gr.215095.116

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag New York. doi: 10.1007/978-0-387-98141-3

Wilson, D. G., Phamluong, K., Lin, W. Y., Barck, K., Carano, R. A., Diehl, L., et al. (2012). Chondroitin sulfate synthase 1 (Chsy1) is required for bone development and digit patterning. *Dev. Biol.* 363, 413–425. doi: 10.1016/j.ydbio.2012.01.005

Wilson, R. T. (1998). *Camels the tropical agriculturalist*. London, United Kingdom: Macmillan Education.

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5, 5188. doi: 10.1038/ncomms6188

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* doi: 10.1016/j.ajhg.2010.11.011.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and reviewer PO-TW declared their involvement as co-editors in the Research Topic.

Copyright © 2019 Bahbahani, Musa, Wragg, Shuiep, Almathen and Hanotte. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 17 October 2019
doi: 10.3389/fgene.2019.00997

The Camel Adaptive Immune Receptors Repertoire as a Singular Example of Structural and Functional Genomics

Salvatrice Ciccarese[1], Pamela A. Burger[2], Elena Ciani[3], Vito Castelli[1], Giovanna Linguiti[1], Martin Plasil[4,5], Serafina Massari[6], Petr Horin[4,5] and Rachele Antonacci[1]*

[1] Department of Biology, University of Bari "Aldo Moro," Bari, Italy, [2] Research Institute of Wildlife Ecology, Vetmeduni Vienna, Vienna, Austria, [3] Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari "Aldo Moro," Bari, Italy, [4] Department of Animal Genetics, Faculty of Veterinary Medicine, University of Veterinary and Pharmaceutical Sciences, Brno, Czechia, [5] CEITEC-VFU, University of Veterinary and Pharmaceutical Sciences, RG Animal Immunogenomics, Brno, Czechia, [6] Department of Biological and Environmental Science and Technologies, University of Salento, Lecce, Italy

The adaptive immune receptors repertoire is highly plastic, with its ability to produce antigen-binding molecules and select those with high affinity for their antigen. Species have developed diverse genetic and structural strategies to create their respective repertoires required for their survival in the different environments. Camelids, until now, considered as a case of evolutionary innovation because of their only heavy-chain antibodies, represent a new mammalian model particularly useful for understanding the role of diversity in the immune system function. Here, we review the structural and functional characteristics and the current status of the genomic organization of camel immunoglobulins (IG) or antibodies, α/β and γ/δ T cell receptors (TR), and major histocompatibility complex (MHC). In camelid humoral response, in addition to the conventional antibodies, there are IG with "only-heavy-chain" (no light chain, and two identical heavy gamma chains lacking CH1 and with a VH domain designated as VHH). The unique features of these VHH offer advantages in biotechnology and for clinical applications. The TRG and TRD rearranged variable domains of *Camelus dromedarius* (Arabian camel) display somatic hypermutation (SHM), increasing the intrinsic structural stability in the γ/δ heterodimer and influencing the affinity maturation to a given antigen similar to immunoglobulin genes. The SHM increases the dromedary γ/δ repertoire diversity. In *Camelus* genus, the general structural organization of the TRB locus is similar to that of the other artiodactyl species, with a pool of *TRBV* genes positioned at the 5' end of three in tandem D-J-C clusters, followed by a single *TRBV* gene with an inverted transcriptional orientation located at the 3' end. At the difference of TRG and TRD, the diversity of the TRB variable domains is not shaped by SHM and depends from the classical combinatorial and junctional diversity. The MHC locus is located on chromosome 20 in *Camelus dromedarius*. Cytogenetic and comparative whole genome analyses revealed the order of the three major regions "Centromere-ClassII-ClassIII-ClassI". Unexpectedly low extent of polymorphisms and haplotypes was observed in all Old World camels despite different geographic origins.

Keywords: Immunome, Old World camelids, *Camelus bactrianus*, *Camelus dromedarius*, *Camelus ferus*, Immunoglobulins, T cell receptors, major histocompatibility complex

OPEN ACCESS

Edited by:
James J. Cai,
Texas A&M University, United States

Reviewed by:
David Nicholas Olivieri,
Universidad de Vgio, Spain
Véronique Giudicelli,
IMGT, the international ImMuno GeneTics information system®,
France

****Correspondence:***
Salvatrice Ciccarese
salvatricemaria.ciccarese@uniba.it

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *21 December 2018*
Accepted: *18 September 2019*
Published: *17 October 2019*

Citation:
Ciccarese S, Burger PA, Ciani E, Castelli V, Linguiti G, Plasil M, Massari S, Horin P and Antonacci R (2019) The Camel Adaptive Immune Receptors Repertoire as a Singular Example of Structural and Functional Genomics. Front. Genet. 10:997. doi: 10.3389/fgene.2019.00997

INTRODUCTION

In vertebrates, B and T lymphocytes together with the antigen-presenting cells play central roles in the adaptive immune system. They respond to a large variety of antigens that are specifically recognized through highly specialized proteins: immunoglobulins (IG) or antibodies in B cells, and T cell receptors (TR) in T cells (Lefranc, 2014a). The common shape of IG is the tetrameric structure, two identical dimers each made up of an IG heavy (H) chain, and an IG light (L) chain (either IG light kappa (IGK) or IG light lambda (IGL) chains). The TR is a heterodimeric receptor that may occur in two types: αβ (TR-Alpha_Beta, composed of a T cell receptor alpha (TRA) and a T cell receptor beta (TRB) chain) and γδ (TR-Gamma_Delta, composed of a T cell receptor gamma (TRG) and a T cell receptor delta (TRD) chain).

Each chain contains a variable domain and a constant region (Lefranc, 2014a). The variable domain forms the antigen-binding site and it is generated during B or T lymphocyte development by a sequential gene rearrangement at the DNA level of the *variable (V)* and *joining (J)* genes of the IGK or IGL, TRG, and TRA loci, and *V*, *diversity (D)*, and *J* genes of the IGH, TRB and TRD loci. After transcription, the rearranged V-(D)-J sequence is spliced to the *constant (C)* gene (Lefranc and Lefranc, 2001; Lefranc and Lefranc, 2001b; Jung and Alt, 2004). The resulting IG and TR chains are proteins with a variable (V) domain at the N-terminal end. Each V domain comprises nine beta sheets forming four framework regions or FR, which support three hypervariable loops (complementarity determining regions or CDR) (Lefranc 2014; Lefranc and Lefranc 2019). CDR1 and CDR2 are encoded by the germline V gene; the third, CDR3, results from the V-(D)-J rearrangement. The six CDR loops of the paired V domains of the IG (VH and VL) and those of the TR gamma/delta (V-gamma and V-delta) contribute to the antigen-binding site. In contrast in the TR alpha/beta, only the two CDR3 principally recognize and bind the antigenic peptide bound to major histocompatibility (MH) proteins of class I (MH1) or class II (MH2), whereas the germline-encoded CDR1 and CDR2 loops mainly contact the helices of the MH proteins (Lefranc, 2014a).

For IGH chains, the rearranged variable domain VH will initially be expressed together with IGHM, the most J-proximal *IGHC* gene, leading to the IgM class synthesis. After the encounter with the antigen and B cell activation and with T cell cooperation, a further DNA recombination event, referred to as class switch recombination, can take place in B cells, resulting in replacement of the IGHM by one of the gene of the other IGHC gene subgroups, namely, IGHG, IGHE, or IGHA. This process leads to the expression of a new H chain with different effector functions, thereby shifting the IG from the IgM class to one of the IgG or IgA subclasses or to IgE class (Lefranc and Lefranc, 2001)

The genes encoded for each IG or TR chain are located in different loci. There are three IG loci (IGK, IGL, and IGH) and four TR loci (TRA, TRB, TRG, and TRD) (http://www.imgt.org/IMGTrepertoire/LocusGenes). The TRA and TRD loci occupy the same chromosome location, being the TRD inserted into the TRA locus. The number of the *V*, *D*, and *J* genes within loci as well as their genomic organization can vary significantly among species. This implies that the gene content is an important element in generating the full extent of the IG and TR repertoires, providing the species with the ability to adapt to its own habitat to defend against infections from a large variety of pathogens.

The complex response of camelids to different pathogens has been investigated over nearly three decades. In this focused review, we provide a comprehensive overview based on the search of key publications from the more recent literature on the genomic and functional characteristics of the IG, TR, and MH molecules in camelids.

THE CAMEL IMMUNOGLOBULIN: A DICHOTOMOUS ADAPTIVE HUMORAL IMMUNE SYSTEM

The humoral immune system of camelids (i.e., *Camelus bactrianus* (Bactrian camel), *C. dromedarius* (Arabian camel), *C. ferus* (Wild Bactrian camel), *Lama glama* (llama), *L. guanaco* (guanaco), *Vicugna pacos* (alpaca), and *V. vicugna* (vicugna) have the particularity of including, in addition to the conventional tetrameric IgG (IgG1 subclass) composed of two identical heavy (H) and two identical light (L) chains connected by disulphide bonds, functional homodimeric IgG (IgG2 and IgG3 subclasses) lacking L chains and, therefore, comprising only two identical H chains (only-heavy-chain antibodies hcAb, or hcIG) (**Figure 1**) (Hamers-Casterman et al., 1993). Sequence and structure analysis revealed a number of characteristic features of camelid hcAb (Muyldermans and Lauwereys, 1999) to make the H chain functional in antigen binding in absence of the L chain. Besides their dissimilar IG chain content, tetrameric, and homodimeric IgG display distinct H chains, with the H chain within hcAb composed of three instead of four globular domains. Biochemical and cDNA sequence analyses have shown that the C region (CHH) of the homodimeric IgG lacks the first domain, CH1, which normally binds to the L chain. This region is eliminated by splicing during mRNA processing due to a point mutation on the donor-splicing site present in the first C exon/intron boundary (Nguyen et al., 1999; Woolven et al., 1999). Hence, the variable domain is joined directly to the hinge region in hcAb (Lefranc, 2014b) (**Figure 1**). The hinge region itself can be different from conventional IG heavy chains (Henry et al., 2019)

Structural Features and Binding Properties of the VHH Domain

Since the hcAb do not contain L chains, the antigen-binding site is reduced to a single domain (referred to as VHH) that resembles the structure of the H chain variable domain (VH) of the tetrameric IgG. However, the VHH domains display remarkable amino acid differences in positions that are conserved in the conventional VH domains. In FR2-IMGT, amino acids highly conserved across species, located at positions 42, 49, 50, and 52, according to the IMGT unique numbering (Lefranc et al., 2003; Lefranc, 2011a; Lefranc, 2011b), which, in conventional VH domains, form the hydrophobic surface associating with VL (Chothia et al., 1985), are changed to more hydrophilic amino acids (Hamers-Casterman et al., 1993; Muyldermans et al., 1994;

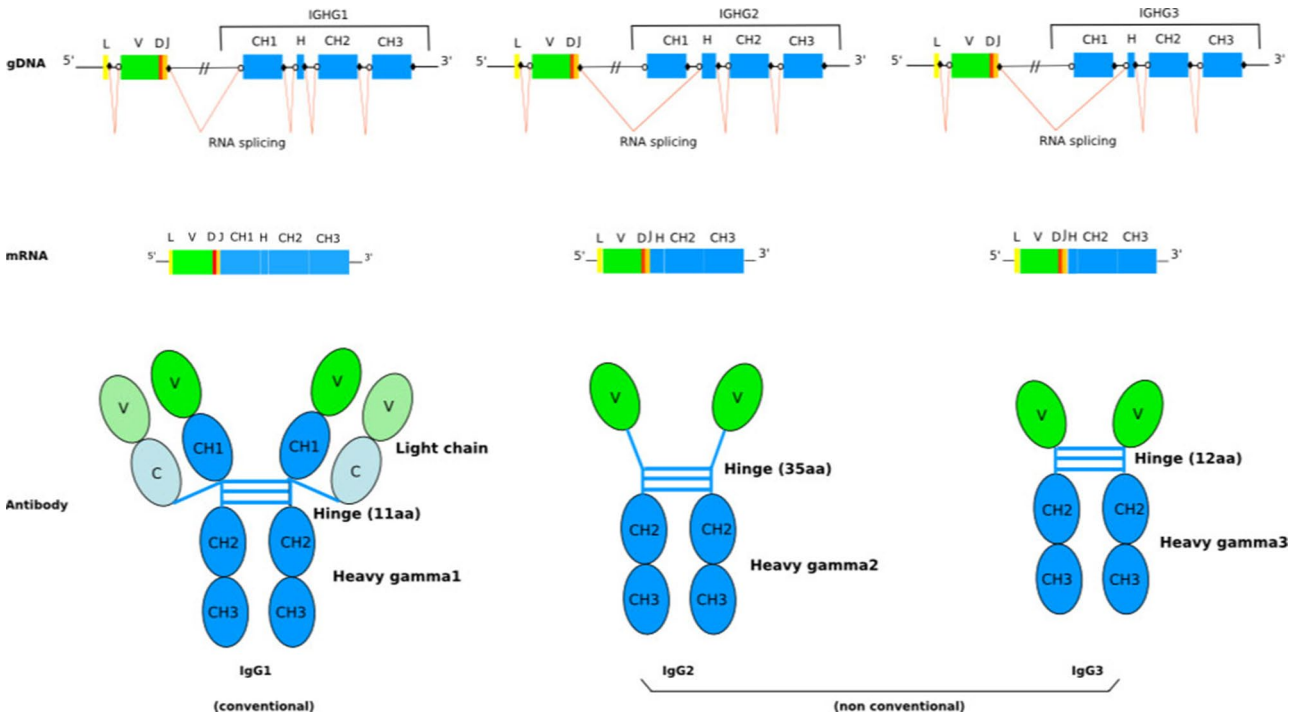


FIGURE 1 | Schematic representation of the general structure of camel conventional IgG1 and the nonconventional IgG2 and IgG3. The lack of light chain is itself a consequence of the absence of the CH1 domain in the gamma 2 and gamma 3 chains, due to a splicing defect of the CH1 exon. The figure is from IMGT: Camelidae IgG antibodies (http://www.imgt.org/IMGTbiotechnology/Camel_IgG.html#characteristics) with permission of IMGT®, the international ImMunoGeneTics information system ® (http://www.imgt.org).

Vu et al., 1997; Tillib et al., 2014; Li et al., 2016; Brooks et al., 2018). The hydrophobic to hydrophilic amino acid changes at these positions make impossible the association of the VHH with a conventional VL domain and undoubtedly help in the solubility behaviour of the hcAb (Davies and Riechmann, 1994; Muyldermans et al., 1994). Moreover, the amino acid change, in most of the Bactrian and Arabian camel V-REGION, of Leu (L) 12 (IMGT numbering) > Ser (S) in FR1, which is seen as an adaptation to the absence of the CH1 domain, helps the solubility of the hcAb. However, the VHH are notably more conserved in sequence and structure across their FR regions than conventional VH domains (Mitchell and Colwell, 2018b).

CDR1-IMGT and CDR2-IMGT of the IGHV encoding VH or VHH domains are highly similar in term of sequences and lengths (IMGT Repertoire (IG and 1TR) > Protein displays Arabian camel (*Camelus dromedarius*)).[1]

Conversely, the CDR3 loop of the hcAb tends to be longer and more variable than that of the conventional IgG (Muyldermans, 2013; Mitchell and Colwell, 2018a; Mitchell and Colwell, 2018b). This difference is much higher in *Camelus bactrianus* than that reported for llama and dromedary (Henry et al., 2019). A longer CDR3 in the hcAb could potentially greatly increase both the sequence and shape diversity of the paratope. As recently reported (Mitchell and Colwell, 2018a), the CDR3 loop is more frequently in contact with the antigen than CDR1 and CDR2 loops. Therefore, CDR3 plays a more dominant role in determining interaction specificity. Moreover, in Bactrian as well as in dromedary camel, the long rearranged CDR3 of VHH in most cases harbours a cysteine that forms a disulphide bond with another additional cysteine located either in CDR1, or on position 50 (*Camelus dromedarius* (Camdro) IGHV) or 55 (*Lama glama* (Lamgla) IGHV) in FR2-IMGT (IMGT Repertoire > IMGT Protein display)[1] (Hamers-Casterman et al., 1993; Muyldermans et al., 1994; Vu et al., 1997). This second disulphide bond stabilizes the VHH domain and fixes the long CDR3 loop into an optimal conformation, increasing the affinity for the antigen (Govaert et al., 2012).

The VHH domain of the hcAb is fully capable of antigen binding. It recognizes a broad range of epitopes with high affinity. Moreover, the hcAb repertoire, largely diversified by extensive somatic hypermutation (SHM) involving the variable domains, results in novel and unusual paratopes different from those of conventional IgG (Nguyen et al., 2000; Nguyen et al., 2001). Many VHH domains are competitive enzyme inhibitors since they interact specifically with the active site of enzymes that, in general, is of low antigenicity for the conventional VH-VL domains (Lauwereys et al., 1998).

The unique characteristics of the VHH and their straightforward bacterial expression have made them of particular interest in biotechnological and pharmaceutical applications. In recent years, the industrialization of camel VHH domains (designated as single-domain antibodies or "nanobodies" for their format) has produced a great expansion of their use

[1] http://www.imgt.org

(Muyldermans et al., 2009; Hassanzadeh-Ghassabeh et al., 2013; Helma et al., 2015; Fernandes, 2018), and in 2018, a diabody of two VHH humanized from *Lama glama*, caplaximab, an anti-von Willebrand factor (VWF) A1 domain, has been approved by the European Medicine Agency (EMA) in Europe and by the Food and Drug Administration (FDA) in 2019 for treatment of acquired thrombotic thrombocytopenia purpura (TTP) [IMGT/mAb-DB, (Lefranc et al., 2015)]. Beyond their application as therapeutics to treat human diseases (Muyldermans, 2013; Rissiek et al., 2014; Steeland et al., 2016), nanobodies have become a valuable research tool. For example, they are used as affinity reagents to assist the crystallization process, to detect antigen trafficking inside living cells, to interfere with protein–protein interactions, and to direct proteins to degradation (Loris et al., 2003; Hassanzadeh-Ghassabeh et al., 2013; Helma et al., 2015; Beghein and Gettemans, 2017; Baudisch et al., 2018; Schumacher et al., 2018).

The Camel IGH Locus

To understand the molecular mechanisms governing the formation of tetrameric and homodimeric IgG in camelids, the characterization of the organization of the genes that encode them is an essential step. Although hcAb have been extensively investigated, to date, there has not been a comprehensive analysis of the repertoire based on high-throughput sequencing (Jirimutu et al., 2012; Li et al., 2016; Ali et al., 2019; Henry et al., 2019), but most efforts have been based upon low throughput sequence analysis, and the reports trying to analyse and describe the complete immune repertoire of camel hcAb are limited. Although the annotated data in public databases are limited, the available sequences show that the camelid *IGHV* genes, which encode VH and VHH, belong to the IGHV3 subgroup (IMGT Repertoire (IG and TR) 2. Proteins and alleles > Protein displays Arabian camel (*Camelus dromedarius*); ibid:alpaca (*Vicugna pacos*) IGHV; and ibid:llama (*Lama glama*) IGHV)[1]. The high percentage of identity between IGHV encoding VH or VHH classifies them in the same IGHV3 subgroup, the differences between them being the characteristic amino acid changes at the four IMGT positions 42, 49, 50, and 52. It is, therefore, the IGHV, which is involved in the rearrangement which determines if the expressed domain is VH or VHH. The constant region of the *Camelus dromedarius* H-gamma1 chains is encoded by the *IGHG1* gene, whereas the constant region of the H-gamma2 and H-gamma3 chains are encoded by *IGHG2* and *IGHG3* genes, which both have a splicing defect of the CH1 DONOR-SPLICE leading to the absence of the CH1 in the transcript, although this CH1 sequence is present in the genomic DNA (**Figure 1**) (Lefranc, 2014b) (IMGT Biotechnology > Antibody camelization > Characteristics of the camelidae (camel, llama) antibody synthesis)[1]. The presence of a point mutation (G to A) in the putative donor splicing site flanking the first C exon and the specific hinge region makes it possible to distinguish the *IGHG* genes encoding the constant region of hcAb chain and the *IGHG* genes encoding the constant region of conventional antibody heavy chain (Vu et al., 1997; Nguyen et al., 1999; Woolven et al., 1999). In Arabian camel, three *IGHG* genes have been identified (IMGT Repertoire (IG and TR) > Gene table > Gene table: Arabian camel (*Camelus dromedarius*) IGHC)[1]. Four *IGHG* genes (*IGHG1A*, *IGHG1B*, *IGHG2B,* and *IGHG2C*, these last two genes without CH1 in the alpaca genome) are present in the alpaca and llama genomes (Achour et al., 2008; Henry et al., 2019) (IMGT Repertoire (IG and TR) > Gene table > Gene table: Alpaca (Vicugna pacos) IGHC)[1], even if other studies have also identified *IGHG2A* and *IGHG3* genes in the expressed IgG repertoire of llama (De Genst et al., 2006; Saccodossi et al., 2012).

IGHV genes encoding VH and VHH domains as well as the *IGHG* genes encoding the C-REGION of conventional IG (IGHG1) and only-heavy-chain IG (IGHG2 and IGHG3) are in an intermixed conformation, in V and C clusters, respectively, in a single IGH locus (**Figure 1**). *IGHV* genes encoding VH and VHH domains recombine with the same *IGHD* and *IGHJ* genes. The identification of identical *IGHD* and *IGHJ* genes in VH and VHH cDNAs suggests the common use of the *IGHD* and *IGHJ* genes (Harmsen et al., 2000; Nguyen et al., 2000; Achour et al., 2008). The *IGHV* genes are designed in three or four subgroups based on their degree of homology with human IGHV subgroups. There are IGHV subgroups that contain exclusively classical *IGHV* genes, whereas one subgroup, IGHV3, contains both classical *IGHV* genes and *IGHV* genes with the FR2-IMGT camelid hydrophilic amino acid in 50 (together with 42, 49, and/or 52) (Lefranc and Lefranc, 2019), which are rearranged and expressed in VH and VHH domains, respectively (Harmsen et al., 2000; Achour et al., 2008; Deschacht et al., 2010; Griffin et al., 2014). The transcription of a VHH domain (V-D-J-REGION) with the IGHG2 or IGHG3 constant region leads to the synthesis of a heavy chain without CH1, which is the characteristic feature of the hcAb repertoire (Hamers-Casterman et al., 1993) (**Figure 1**). However, it has been reported that also classical *IGHV* genes can contribute to the hcAb pool. In this case, the V domain has a classical short junction without the additional cysteine (compared to the VHH) and the conserved anchor Trp 118 is substituted mostly by an arginine codon (Deschacht et al., 2010). It was shown that also the IGHV4 subgroup contributes to produce both classical IgG and hcAb. Interestingly, a same IGHV4-IGHD-IGHJ rearrangement has been shown to be shared between a classical tetrameric IgG and a dimeric hcAb (heavy chain with no CH1, and no light chain).

From a limited number of germline *VHH* genes, camelids can generate a large and diversified repertoire by extensive SHM (Nguyen et al., 2000). The pattern of variability particularly within CDR loops but also in framework regions is larger in VHH than in VH cDNAs. As for classical VH, crystallographic studies of the VHH-antigen complexes demonstrated that amino acids located in the CDR1 and CDR2 loops and, particularly, the long CDR3 interact with the antigen (Desmyter et al., 1996; Decanniere et al., 1999). Contact analysis between the VHH and the ligand is provided in IMGT/3Dstructure-DB, in 38 entries of *Camelus dromedarius* and 58 for *Lama glama* (August 2019). Thus, the introductions of mutations together with a long CDR3 increase the VHH potential repertoire for antigen binding.

DROMEDARY *TRG* AND *TRD* GENES

The Genomic Organization of the Dromedary TRG Locus

γδ T cells have unique features when compared with the more abundant αβ T cells, e.g., a preferential distribution in both epithelial and mucosal sites, and an immunoglobulin like antigen recognition mechanism in addition to the MH-restricted one. In the immune response during inflammatory processes, γδ T cells release cytokines and kill infected macrophages; they combine the characteristics of an innate-like immune response with those of an adaptive response to inflammation (Allison et al., 2001; Allison and Garboczi, 2002; Adams et al., 2005). Their percentage in peripheral blood cells, depending on age and species, differs strikingly from that of αβ T cells (Carding and Egan, 2002). Artiodactyls (sheep, cows and pigs) are referred to as "γδ-high species" since they exhibit a higher frequency and a wider physiological distribution of γδ T cells with respect to other mammalian species, including humans and mice, which are referred to as "γδ-low species" (Hein and Dudler, 1993; Ciccarese et al., 1997).

Recent studies have shown the presence of SHM in the γ chain of gamma/delta in shark T cells and in both chains γ and δ of the dromedary camel (Chen et al., 2009; Antonacci et al., 2011; Chen et al., 2012; Vaccarelli et al., 2012). In each work, it was shown that SHM followed the characteristics of the mutational profiles detected in the B cells that undergo affinity maturation. γδ T cells, unlike αβ T cells, interact with nonclassical major histocompatibility complex (MHC) and have a small number of genes, so they have a limited diversity (Allison et al., 2001; Adams et al., 2005). The SHM seems to be used as a mechanism for further diversification of the γδ receptor and for an optimal recognition of the ligand. Consequently, this would allow the evolutionary changes in the loci, which, in turn, allow the receptor to evolve more rapidly in mutant environments (Adams et al., 2005; Kazen and Adams, 2011).

In "γ/δ-high" species, the TRG and TRD expressed repertoire is mainly affected by a large number of genes distributed in reiterated duplications of functional TRG cassettes (Vaccarelli et al., 2005; Conrad et al., 2007; Vaccarelli et al., 2008) and by a marked expansion and preferential usage of the TRDV1 multigene subgroup (Ishiguro et al., 1993; Yang et al., 1995; Hein and Dudler, 1997; Massari et al., 2000; Antonacci et al., 2005). Usually, in mammals, less than a few exceptions such as in human (Lefranc and Rabbitts, 1989) and in dolphin (Linguiti et al., 2016), TRG loci are quite complicated, containing numerous *V*, *J*, and *C* genes, sometimes located in different chromosomal bands (Massari et al., 1998; Miccoli et al., 2003), or spanning hundreds of kb (Massari et al., 2009).

In *Camelus dromedarius,* the TRG locus spans approximately 45 kb and it maps in a homology region established between bovids chromosome 4, human chromosome 7, and pig chromosome 9, where orthologous TRG loci have been mapped (Vaccarelli et al., 2012).

The dromedary locus consists of two *TRGV* genes (*TRGV1* and *TRGV2*), four *TRGJ* genes, and two *TRGC* genes (*TRGC1* and *TRGC2*), all in the same transcriptional orientation, organized in two functional cassettes (5'-TRGV1-TRGJ1-TRGJ1-2-TRGC1 and TRGV2-TRGJ2-1-TRGJ2-2-TRGC2 -3'). Considering the exon organization of the ovine and human C regions, we inferred that both the dromedary C regions keep a connecting region encoded by three different exons, as is observed in the sheep *TRGC2*, *TRGC4*, and *TRGC6* genes (Miccoli et al., 2001) and in the polymorphic human *TRGC2* gene (Lefranc and Lefranc, 2001b). The dromedary locus organization in two (V-J-J-C) cassettes potentially limits the combinatorial usage of its genes. However, cDNA sequencing clearly revealed that, besides the combinatorial diversity and the introduction of N region diversity typical of all known *IG* and *TR* genes (Lefranc and Lefranc, 2001a; Lefranc and Lefranc, 2001b), the SHM mechanism enhances the TRG and TRD repertoire diversity in *Camelus dromedarius* (Ciccarese et al., 2014).

SHM in TRG and TRD V Domains and Nature of AA Changes

Among mammals, SHM occurs primarily in germinal center B cells, it introduces point mutations into the variable domains of IG, and it is the driving force for antibody affinity maturation (Li et al., 2004). During SHM, the dgyw/wrch motif (where d = a or g or t, y = c or t, w = t or a, r = a or g, and h = c or t or a) has been found to be the principal hotspot for activation cytidine deaminase (AID) inducing g:u lesions in rearranged *IG* genes (Rogozin and Diaz, 2004; Liu and Schatz, 2009). These changes are dominated by point mutations and biased toward transitions (G:A and C:T). Moreover, the principal site for a/t mutations has been identified in the dinucleotide target wa/tw. This secondary mutator has allowed to define the roles of the error-prone polymerases in mismatch repair (Pavlov et al., 2002; Zhao et al., 2013).

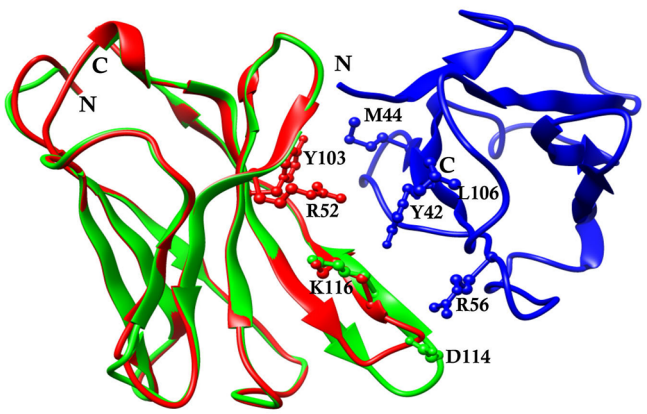
The features of mutations were evaluated comparing the genomic sequence of the single *TRDV4* gene with the TRDV4 cDNA clones derived from spleen and blood, excluding CDR3 and including the TRDJ4 genomic sequence. Four related sets of TRDV4 clones with common CDR3 sequences and unrelated sequences derived from independent rearrangements in blood and spleen were deduced (Antonacci et al., 2011). The analysis of mutations identified a clonal genealogy among related sequences, showing that tandem mutations resulted from sequential point mutations and, given the TRDV4 gene uniqueness, the nucleotide substitutions are not the result of gene conversion events (Antonacci et al., 2011). AID-dependent deamination of cytidine to uracil produces mutations at c/g nucleotides activates the repair proteins MSH2-MSH6 to bind u:g mismatches and recruits the lowfidelity DNA polimerase η (Pol η) (Wilson et al., 2005; Schanz et al., 2009; Zhao et al., 2013). The analysis of substitutions in the TRDV1 and TRDV4 mutated sequences show a bias for transition changes in blood and spleen. Therefore, the nature of V domain resulting from dromedary *TRG* and *TRD* rearranged genes is the result of a combined action of AID, uracil-DNA glycosylase (UNG), and mismatch (MSH) repair pathways (**Supplementary Figure 1**) (Vaccarelli et al.2012; Ciccarese et al., 2014).

In mutated TRGV1 and TRGV2 clones, replacement mutations occur preferentially in bases inside the [dgyw] and [wrch] AID motifs, whereas neutral mutations are favoured

outside the motifs. Chi-square analysis used to compare observed and expected numbers of replacement mutations between FR and CDR regions highlighted no significant difference in TRGV1 and in TRDV4 clones. On the contrary, in TRGV2 and TRDV1 clones, the difference in the R/S ratio between CDR and FRs was significant. This feature is consistent with a selection pressure acting on dromedary γδ T cells and with the affinity maturation during clonal expansion.

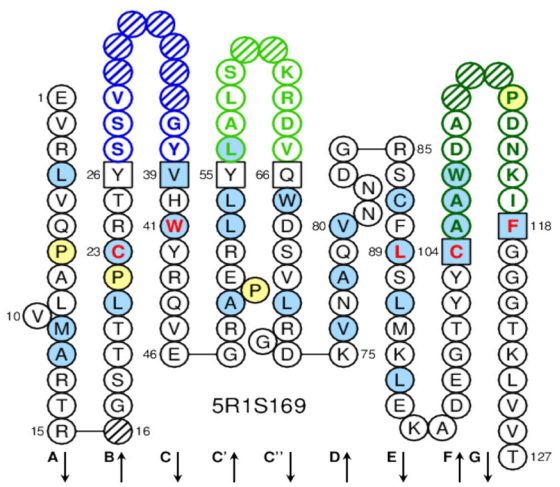
When shared mutations were used as position set-points to construct putative lineages from TRG and TRD cDNA clone sets, each of them sharing an identical CDR3, the inferred lineages harbouring multiple radiations highlighted that the substitutions that arose first (progenitor mutations) were replacements changes. These progenitor mutations occurred starting from the CDR3/FR4 region and proceeding toward the leader region along the variable domain with additional mutations in FR2 and FR3. Moreover, it was found that progenitor mutations are selected and transmitted during clonal expansion and they are nonconservative of the amino acid physicochemical properties, i.e., change nonhydrophobic amino acid residues to hydrophobic ones. A computed model was constructed with the TRGV2 translated sequences of the corresponding mutated cDNAs; these sequences were visualized in their two-dimensional structure with the IMGT tool Collier de Perles (Ehrenmann et al., 2011; Lefranc, 2011a; Lefranc, 2011b). The comparative modelling procedure was applied using the counterpart γδ human T cell receptor subunits (Ehrenmann et al., 2010; Xu et al, 2011) and a notable difference between the genomic paired TRGV2/TRDV4 and the mutated cDNA TRGV2 paired to genomic TRDV4 was observed (**Figure 2**). Only for the last paired V domains, the occurrence of putative hydrogen bonds and salt bridges confirmed that the changes alter the conformation of the variable domains of the γδ receptor with consequent effects on its stability.

If the replacements of hydrophilic amino acid residues with hydrophobic ones are maintained and positively selected during the proliferation of T cells, it follows that they stabilize the structure of the receptor whether they fall into the CDR or into the FR. Therefore, conclusions are consistent both with the acquisition of new antigenic specificities and repertoire diversification and simultaneously with selection for changes in paratopes in the manner similar to that of immunoglobulin gene during the B cells affinity maturation to a given antigen. The same conclusions were recently reached by Ott et al., in a paper where the authors propose that the SHM in TRA chain contributes to selection of αβ T cells in nurse "couch potato" shark



TRDV4	Dist. [Å]	TRGV2
Y42 [OH]	3.14	K116 [O]
M44 [SD]	3.82	Y103 [OH]
L106 [O]	3.74	R52 [NH2]

TRDV4	Dist. [Å]	5R1S169
Y42 [OH]	3.00	K116 [O]
Salt Bridges		
R56 [NH1]	3.69	D114 [OD1]
R56 [NH1]	3.63	D114 [OD2]




```
TRGV2      EVRLVQPALVMARTRGSTTLPCRTYGSVTYVHWYR
5R1S169    EVRLVQPALVMARTRGSTTLPCRTYSSVGYVHWYR
           ************************** ** ******

TRGV2   QLEGRAPERLLYLALSKRDVQWDSVLRGDKVNAQVNNDGR
5R1S169 QVEGRAPERLLYLALSKRDVQWDSVLRGDKVNAQVNNDGR
        * ********** ********** ********************

TRGV2   SCFLSLMKLEKADEGTYYCAAWDAELYNKIFGGGTKLVVT
5R1S169 SCFLSLMKLEKADEGTYYCAAWDAELYNKIFGGGTKLVVT
        ************************************ **********
```

FIGURE 2 | Sequence comparison of dromedary TRGV2/5R1S169, and the computed interaction between the genomic TRGV2, its mutated counterpart 5R1S169, and the genomic TRDV4 (modified from Ciccarese et al., 2014). Structural overlapping of TRGV2 (depicted in red), and 5R1S169 clone (depicted in green) with TRDV4 (depicted in blue), is shown. CDR-IMGT regions are framed with colored rectangles according to IMGT standardized colors. IMGT Collier de Perles of 5R1S169 clone is shown (Lefranc et al., 2003; Lefranc, 2011a; Lefranc, 2011b). The protein complex interface was computed by the online tool PDBePISA at the EBI server. (http://www.ebi.ac.uk/msd-srv/prot_int/) and visualized by UCSF Chimera tool (http://www.cgl.ucsf.edu/chimera/).

thymus (Ott et al., 2018). The absence of SHM for the TRB chain reported recently in expression assays of spleen in dromedary (Antonacci et al., 2017a) could constitute the watershed between the cartilaginous fish that are the most divergent jawed vertebrate group relative to mammals, and mammals in the scenario of the thymic selection of αβ and γδ T cells.

In the dromedary TRG and TRD loci, evolution allowed the SHM to increase the receptor repertoire of cell-mediated immunity. Previously, we have proposed that requirements related to immunoprotective functions, including the first defensive barrier in the epithelia of the digestive tract, are likely to have induced in TRG and TRD loci of ruminants a sort of genome functional fluidity resulting in duplications of *TRG* gene cassettes and in a marked expansion of the *TRDV1* multigene subgroup (Vaccarelli et al., 2012).

In this review, we point out that, in dromedary, TRG and TRD evolution was favoured by mutation in the productively rearranged *TRG V-J* and *TRD V-D-J* genes, so that a large and diversified TRG and TRD repertoire could be generated even in absence of functional reiterated gene duplications. Because SHM has not been shown to occur in any mammalian organism, we can hypothesize that Camelidae by themselves might occupy a peculiar immunological niche, proposing the camel lineage as a fascinating model in the evolution of immune systems.

THE ORGANIZATION AND EVOLUTION OF *CAMELUS* TRB LOCUS IS SHARED IN TYLOPODA, SUINA, AND RUMINANTIA

The organization of the TRB locus has been extensively investigated in different mammalian species and it consists of a general structure with a group of *TRBV* genes located at the 5' end of the locus followed by in tandem TRB D-J-C clusters. A *TRBV* gene, with an inverted transcriptional orientation, lies at the 3' end of the region. A common aspect to most species, such as humans, rabbits, and dogs (Mineccia et al., 2012; Antonacci et al., 2014; Lefranc et al., 2015), is the presence of two TRB D-J-C clusters, each composed of one *TRBD*, several *TRBJ*, and one *TRBC* genes. Instead, three TRB D-J-C clusters composed the TRB locus in artiodactyl species, i.e., sheep, cattle, and pig (Antonacci et al., 2008; Connelley et al., 2009; Eguchi-Ogawa et al., 2009; Massari et al., 2018). The additional TRB D-J-C cluster 3 is located between the conserved TRB D-J-C cluster 1 and 2 (**Figure 3**). The sequence analysis revealed that the new TRB D-J cluster is correlated to the last one, whereas the *TRBC3* gene is more similar to the *TRBC2* gene in the first part and to the *TRBC1* gene sequence in the last part.

The structure of the TRB locus has been recently investigated in *Camelus dromedarius* (Antonacci et al., 2017a; Antonacci et al., 2017b) and in its wild and domestic Bactrian camel congeners, *Camelus ferus* and *Camelus bactrianus* (Antonacci et al., 2019). The analysis showed that the camel TRB organization is similar to that of the other artiodactyl species, with the presence of three TRB D-J-C clusters (**Figure 3**). This outcome suggests that the TRB genomic organization with three TRB D-J-C clusters was established, prior the Tylopoda/Ruminantia/Suina divergence, through a duplication event due to an unequal crossing-over between the ancestral *TRBC1* and *TRBC2* genes. Following duplication, subsequent species-specific diversifications were made that led to the current genomic organization of the 3' end of the TRB locus in the different artiodactyl species.

As in all mammalian species, *MOXD2* and *EPHB6* genes border the camel locus at the 5' and 3' end, respectively, whereas *TRY* genes are interspersed among *TRBV* genes and arranged in two distinct genomic positions (Antonacci et al., 2017a; Antonacci et al., 2017b; Antonacci et al., 2019).

An expression assay conducted on dromedary spleen T cells (Antonacci et al., 2017a) has demonstrated that all the three TRBD-J-C clusters are used to generate a functional TRβ chain increasing the combinational and junctional diversity of the CDR3 domain. Moreover, the analysis of the cDNA collection shows a preferential usage of the *TRBD1* gene followed by the TRBD3 and TRBD2. This may result in a greater efficiency of the PDβ1 promoter with respect to the PDβ3 and PDβ2, whereas the activity of two similar PDβ3 and PDβ2 could be correlated with their position from 5' to 3' within the locus. Furthermore, a prominent utilization of the *TRBJ3* gene set with respect to the TRBJ2 and TRBJ1 clusters was also observed, probably depending on the number of genes that lie in the genomic region. Probably, multiple 12 bp spacer-recombination signal sequence (12-RS) located in a restricted region may increase the local concentration of the RAG protein that mediates the recombinant process (Di Tommaso et al., 2010). Beside the number of *TRBD* and *TRBJ* genes, other mechanisms seem to increase the dromedary TRB chain functional repertoire, including the incorporation of two *TRBD* genes in the rearrangement process, the intercluster recombination, and the trans-rearrangement (Antonacci et al., 2017a).

While the structure of the TRB D-J-C clusters is similar to the other artiodactyl species, the 5' end of the camel TRB locus appears to be different with a contraction of the total number of the TRBV genes link to a reduction of duplicated events within the TRBV cluster (Antonacci et al., 2017a; Antonacci et al., 2017b; Antonacci et al., 2019). 30 genes in *Camelus ferus* and 33 in *Camelus dromedarius* as in *Camelus bactrianus*, in all cases assigned to 26 different subgroups, are a low number when compared to the 134 *TRBV* genes for bovine, 67 for human and 74 for rabbit, but only slightly lower than that of pig (38 *TRBV* genes) and dog (37 *TRBV* genes) (Connelley et al., 2009; Mineccia et al., 2012; Antonacci et al., 2014; Lefranc et al., 2015; Massari et al., 2018).

The phylogenetic analysis of the *TRBV* genes (**Figure 4**) shows that each of *Camelus ferus*, *Camelus bactrianus*, and *Camelus dromedarius* subgroups come together and form a monophyletic group with a corresponding *TRBV* gene in human and, if present, in dog, sheep, and pig. This is consistent with the occurrence of distinct subgroups prior to the divergence of the different mammalian species. Three human TRBV subgroups (TRBV4, TRBV17, and TRBV18) are lacking in all three camel species, indicating that these subgroups have been lost in these species (i.e., TRBV4) or alternatively they might have originated after the separation of Camelidae (i.e., TRBV18) or Artiodactyla (i.e., TRBV17) from the other mammalian species.

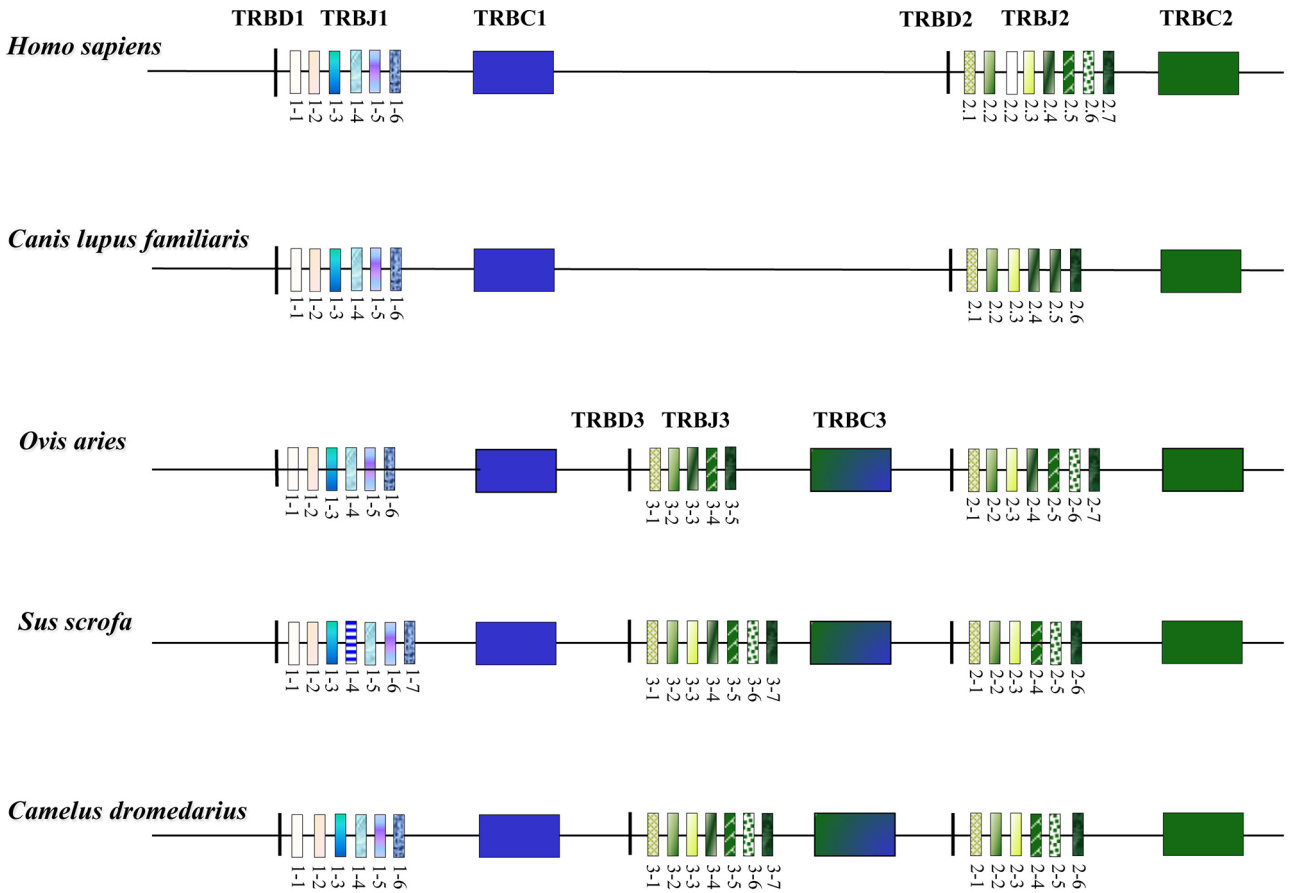


FIGURE 3 | Schematic representation of the D-J-C clusters within the TRB locus of the different mammalian species modified from Antonacci et al., 2017a. The orthologous and paralogous *TRBJ* genes are depicted with the same colour as deduced by the phylogenetic analysis. The double colour *TRBC3* gene indicates the homologies with the *TRBC1* and *TRBC2* genes as result of the unequal crossover that generated the new TRB D-J-C cluster in the artiodactyl species (Antonacci et al., 2008; Eguchi-Ogawa et al., 2009).

THE MHC IN CAMELS

The Overall Organization and Genetic Diversity

The MHC has evolved as part of adaptive immunity in vertebrates. In mammals, it spans approximately 4 Mbp and harbours genes, encoding hundreds of proteins with different immune as well as nonimmune functions. The mammalian MHC locus is a complex genomic region that evolved from an ancestral MHC locus, encoding primarily antigen presenting molecules, which are expressed on the surface of the cell and are involved in the immune system's defence to recognize foreign ("nonself") substances.

Two classes of antigen-presenting molecules and their genes can be distinguished. MH class I molecules present antigenic peptides originating from self as well as nonself (e.g., virus-encoded) intracellular proteins. MH class II proteins present peptides derived from extracellular proteins (e.g., bacterial products), which were internalized by specialized cells of the immune system. While molecules encoded by MH class I and II genes are mainly responsible for antigen presentation to T lymphocytes (adaptive immunity), the MH class III region includes multiple genes involved among others in the innate immune system, such as tumour necrosis factor alpha (*TNFA*) and members of the complement cascade, which help to eliminate invading pathogens (Trowsdale and Knight, 2013).

In Old and New World camelids, the MHC is positioned on the long arm of chromosome 20 (Avila et al., 2014; Plasil et al., 2016). In Old World camels, available genome sequence data (Wu et al., 2014; Fitak et al., 2016a; Fitak et al., 2016b) were combined with initial next generation sequencing to characterize the overall MHC organization (**Figure 5**). The most common general structure of the mammalian MHC region, represented by the order of genes MH class II – MH class III – MH class I, was also confirmed in camels (Plasil et al., 2016; Plasil et al., 2019). Based on the nucleotide sequences retrieved, the MHC of camels seems to be, at least in some of its subregions, more similar to the human and pig MHC rather than to cattle, a phylogenetically closely related species. Besides differences in the location of the *TAP1* gene, phylogenetic analyses of individual MHC genes showed that certain MHC subregions of camels are also closer to pigs than to cattle. Detailed phylogenetic analyses can be found

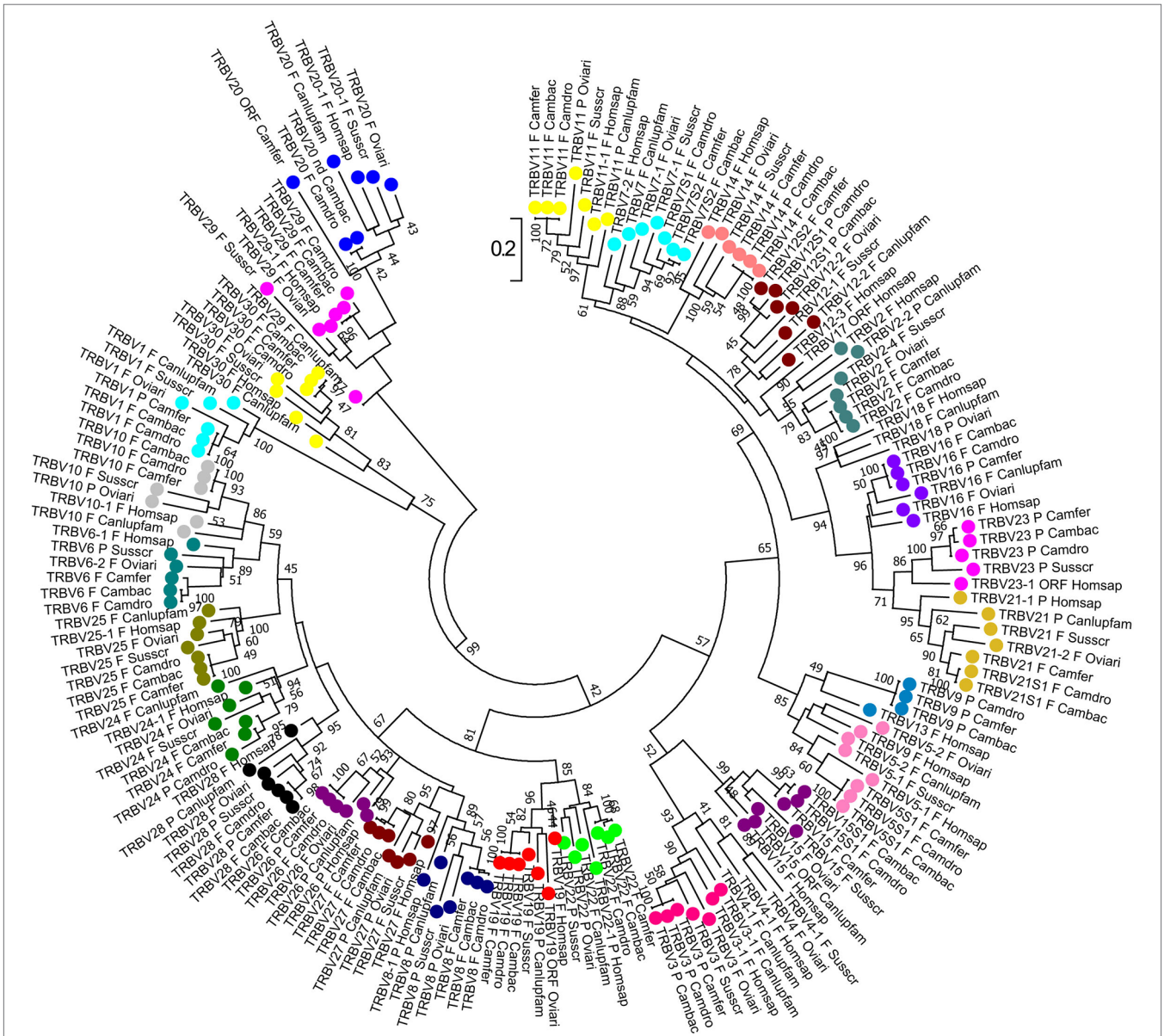


FIGURE 4 | Phylogenetic tree of the dromedary, wild, and domestic Bactrian camel, human, dog, sheep, and pig *TRBV* gene sequences. The sequences of the human, dog, sheep, and pig *TRBV* genes were retrieved from the IMGT database (http://www.imgt.org/), whereas the camel *TRBV* gene sequences derived from Antonacci et al., 2019. The nucleotide sequences of the V-REGION were combined by adopting two selection criteria: (1) only potential functional and in-frame pseudogenes (excepted for human TRBV1) were included; and (2) only one gene for each of the subgroups was selected for each species. The evolutionary analysis was conducted in MEGA7 (Kumar et al., 2016). The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei, 1993). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 181 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. The different colours highlight the distribution of the phylogenetic groups corresponding to the 26 camel TRBV subgroups. The gene functionality according to IMGT rules (F, functional; ORF, open reading frame; P, pseudogene) is indicated; nd indicates that the nucleotide sequence of the gene is incomplete and its functionality cannot be defined. The IMGT 6-letter for species (Homsap, Susscr, Oviari, Camdro, Camfer, and Cambac) and 9-letter for subspecies (Canlupfam) standardized abbreviation for taxon is used.

in Plasil et al., 2016 for MH class II and in Plasil et al., 2019 for MH class I.

MH class I and II molecules are heterodimers consisting of two polypeptidic chains (α and β). While both α and β polypeptides of the MH class II heterodimers are encoded within the MHC, the MH class I β-2-microglobulin is encoded by a gene located on a different chromosome. In camels, it is on chromosome 6. MH class I loci can be distinguished into "classical" and "nonclassical" MH class I genes, the former coding for antigen-presenting molecules, thus being highly polymorphic, whereas the latter

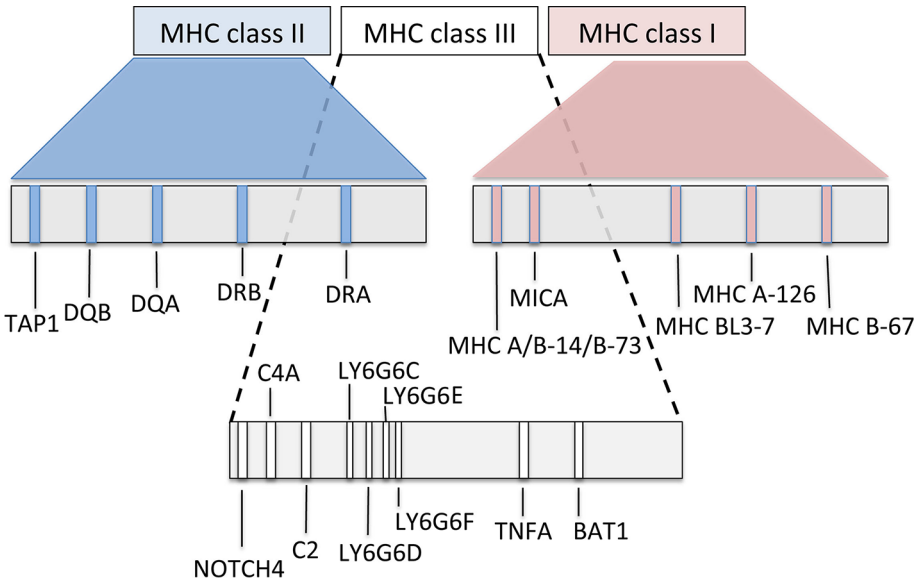


FIGURE 5 | Schematic map of the overall organization of the major histocompatibility complex in Old World camelids, modified from Plasil et al. (2016, 2019).

less polymorphic genes code for a group of structurally related proteins with a variety of immune functions. Some of them may even be considered as part of the innate immune system (Allen and Hogan, 2013). MH class II genes are represented by groups of often duplicated loci (e.g., DR, DB, DM, DO, DP, and DQ). Within such a group, *A* genes (e.g., *DRA*) encode the alpha chain and *B* genes (e.g., *DRB*) the beta chain of the class II heterodimer. Most of functionally relevant polymorphisms are concentrated in the antigen binding sites (ABS) of the molecule, encoded by MH class I exons 2 and 3 and by exon 2 of class II genes.

Overview of the *MHC* Genes Polymorphism

The most characteristic feature of the MHC genes is their high polymorphism, i.e., high numbers (up to hundreds) of allelic variants, especially for genes encoding antigen-presenting molecules. In fact, the MHC is one of the most polymorphic regions in the genome (Janeway et al., 2001). This is important for the immune system to respond fast to rapidly evolving pathogens, a mechanism also described as an evolutionary "arms race" (Sironi et al., 2015).

The genetic diversity of the MHC of camels was first studied in exon 2 sequences of selected class II genes. A surprisingly low level of polymorphism of the *DRA*, *DRB*, and *DQB* genes was observed in all three Old World camel species (Plasil et al., 2016). For the *DRA* locus, *DRA exon 2* spanning 246 bp contains one synonymous and one nonsynonymous single nucleotide polymorphism (SNP) combined in three different alleles shared between dromedaries, domestic and wild Bactrian camels. Successful amplification of this gene in ancient (13^{th} - 16^{th} century, common era) dromedary specimens resulted in three additional substitutions when compared to the reference sequence (Plasil et al., 2016). The *DRB exon 2* (270 bp) contains five polymorphisms shared between the three Old World camel species. In *DQA exon 2,* 11 SNPs, 4 of them synonymous, were identified in *Camelus bactrianus*, of which nine were shared with *Camelus ferus*. In total, three haplotypes were detected, one of them common to all three species and another one shared only between domestic and wild Bactrian camels. The remaining allele was found only in domestic Bactrian camels. The *DQB exon 2* locus was the most polymorphic with 21 polymorphic sites identified across the Old World camels (Plasil et al., 2016). The *DQB* exon 2 harbours a 12-bp long potentially functional insertion not observed in other mammalian species. Since a complete *DQB* exon 2 sequence was not retrieved, the overall extent of polymorphism in this locus remains undefined. However, data available so far suggest that only limited numbers of haplotypes may exist, similarly to the *DQA* locus containing comparable numbers of SNPs. The *DQB* locus is therefore still under investigation.

Similar observations of low diversity were made for MH class I and related loci. In the *classical locus B-67*, only one synonymous polymorphism was found in the entire exon 2 - 3 region. This SNP is shared between dromedaries and Bactrian camels. The *BL3-7* gene is a locus of unclear status, highly similar to the annotated sequence of the *BL3-6* in alpacas (Avila et al., 2014). Interestingly, it is also closely related to the locus *SLA-11* in pig, one of MHC loci with unknown function and unusual structure. In the Old World camels, this gene contains four SNPs (Plasil et al., 2019). The *MH class I related locus MR1* is an antigen-presenting molecule contributing to the regulation of the microbiome in the intestinal tract (McWilliam et al., 2016). Over the total 22 kbp long *MR1* sequence (located on chromosome 21), 170 polymorphic sites were identified, 5 of them located in the coding sequence and partially shared between dromedaries and Bactrian camels (Plasil et al., 2019). The *MH class I related locus MICA* functions as a stress signalling molecule recognized by the NKG2-D type II receptor on natural killer (NK) cells, αβ

T-cells, and γδ T-cells (Shafi et al., 2011; Xiao et al., 2015). A total of 40 SNPs were observed in this sequence, of which eight were found in the coding region (Plasil et al., 2019). It is rather unusual that these MH class I related loci are more polymorphic than a classical MH class I locus, *B-67*. However, knowledge on MH class I classical genes is rather limited and a better-supported conclusion on their diversity can be done only after an extensive analysis of the entire subregion. The first step toward such an analysis and toward an analysis of the MH class III subregion will be made by annotating further genes in a new genome assembly of the dromedary.

CONCLUSION

The Camelidae species occupy an important immunological niche within the humoral as well as cell mediated immune response. In additional to the conventional IG, the serum contains a significant amount of IgG composed solely of paired H chains, which are largely diversified by extensive SHM, resulting in novel paratopes different from those of conventional IgG. The antigen binding fragment of these unique hcAbs comprises only one single domain. When produced by microbial expression system, these recombinant miniature antigen binding fragments possess beneficial biophysical properties useful as research tools and for *in vivo* pharmacological applications as candidate drugs for the treatment of human diseases (Muyldermans, 2013; Rissiek et al., 2014; Steeland et al., 2016).

Moreover, an SHM mechanism in productively rearranged *TRD* and *TRG* genes never identified in mammalian species so far, increases the repertoire diversity of the dromedary γδ T cells that recognize the antigen in a manner antibody like. In this contest, the structural changes within the γδ heterodimer, which is stabilized by mutations both in FR and in CDR in genealogical related clones, could enable the acquisition of new antigenic specificity and, at the same time, could influence the affinity maturation to a given antigen in a manner similar to that of *IG* genes (Ciccarese et al., 2014).

As single-chain antibodies and SHM in *TR* genes were described also in cartilaginous fish (Chen et al., 2009; Flajnik et al., 2011; Chen et al., 2012; Ott et al., 2018). It can be argued that a molecular convergence of the adaptive immune response between these species does exist.

Conversely, in αβ T cells, the limited germline TRBV repertoire in *Camelus dromedarius* as in *Camelus ferus* and *Camelus bactrianus* with a great sequence identity between orthologous genes (Antonacci et al., 2017a; Antonacci et al., 2017b; Antonacci et al., 2019) is not shaped by SHM and it might be related to the constraint imposed on αβ CDR1 and CDR2 domains by the requirements for binding to MH molecules, which, in turn, show a low level of genetic diversity in all three camel species (Plasil et al., 2016).

In Old World camels, the MHC region is structured similarly to a number of other mammalian species. Phylogenetic relationships of camel *MHC* genes do not always follow relationships based on other (neutral) nuclear genes. The diversity of the MH class I, II, and class I – related genes is generally lower than expected. This observation is consistent with a low genome-wide nuclear diversity in dromedaries, wild and domestic Bactrian camels (Fitak et al., 2016a; Fitak et al., 2016b). Several bottlenecks, in the evolutionary history of the three species, but also in recent times due to domestication of dromedaries and Bactrian camels or hunting and habitat decline of the wild camels, respectively, might be responsible for the reduced immunogenetic and genome-wide variability in Old World camels. However, experimental evidence for answering the question whether the low MHC diversity is really due to low diversity of the camel genomes is still lacking.

As the camel is a useful and promising model for therapeutic applications and for phylogenetic and evolution studies about the humoral and cell mediated immunity in jawed vertebrates, implementation of the camelid genomic sequences of IG, TR, and MHC loci is necessary to encourage progress for improvement of the global knowledge of the adaptive immune responses of these animal models.

AUTHOR CONTRIBUTIONS

SC, PB, SM, and RA designed and wrote the review; MP and PH contributed to manuscript writing; EC, VC, and GL contributed to searching of genomic literature. All authors have read and approved the final manuscript.

FUNDING

Austrian Science Fund (FWF): P29623-B25.

ACKNOWLEDGMENTS

We deeply acknowledge the overall support of camel breeders and the wild camel protection foundation for providing samples and in-depth knowledge about camels. PB acknowledges funding from the Austrian Science Fund (FWF): P29623-B25. The financial support of the University of Bari and of the University of Salento is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00997/full#supplementary-material

SUPPLEMENTARY FIGURE 1| Percentages of replacement (R) and silent (S) mutations inside and outside the activation-induced cytidine deaminase (AID) target motifs in TRGV1 and TRGV2 cDNA. Replacement mutations [(R); blue areas of bars] occur preferentially in bases inside the (a/g/t)g(c/t)(a/t) [or dgyw] and (a/t)(a/g)c(c/t/a) [or wrch] AID motifs, whereas neutral mutations [silent (S); red areas of bars] are favoured outside the motifs. Strikingly, in the coding strand of the wrch hot spot, most of the cytidine mutations are replacements (Ciccarese et al., 2014). This figure is reproduced with permission from Elsevier (License number 4515441285009).

REFERENCES

Achour, I., Cavelier, P., Tichit, M., Bouchier, C., Lafaye, P., and Rougeon, F. (2008). Tetrameric and homodimeric camelid IgGs originate from the same IgH locus. *J. Immunol.* 181, 2001–2009. doi: 10.4049/jimmunol.181.3.2001

Adams, E. J., Chien, Y. H., and Garcia, K. C. (2005). Structure of a gammadelta T cell receptor in complex with the nonclassical MHC T22. *Science* 308, 227–231. doi: 10.1126/science.1106885

Ali, A., Baby, B., and Vijayan, R. (2019). From desert to medicine: a review of camel genomics and therapeutic products. *Front. Genet.* 10, 17. doi: 10.3389/fgene.2019.00017

Allen, R. L., and Hogan, L. (2013). "Non-Classical MHC Class I Molecules (MHC-Ib)," in *eLS* (Chichester, UK: John Wiley & Sons Ltd). doi: 10.1002/9780470015902.a0024246

Allison, T. J., and Garboczi, D. N. (2002). Structure of gammadelta T cell receptors and their recognition of non- peptide antigens. *Mol. Immunol.* 8, 1051–1061. doi: 10.1016/S0161-5890(02)00034-2

Allison, T. J., Winter, C. C., Fournié, J. J., Bonneville, M., and Garboczi, D. N. (2001). Structure of a human gammadelta T-cell antigen receptor. *Nature* 411, 820–824. doi: 10.1038/35081115

Antonacci, R., Lanave, C., Del Faro, L., Vaccarelli, G., Ciccarese, S., and Massari, S. (2005). Artiodactyl emergence is accompanied by the birth of an extensive pool of diverse germline TRDV1 genes. *Immunogenetics* 57, 254–266. doi: 10.1007/s00251-005-0773-7

Antonacci, R., Di Tommaso, S., Lanave, C., Cribiu, E. P., Ciccarese, S., and Massari, S. (2008). Organization, structure and evolution of 41 kb of genomic DNA spanning the D-J-C region of the sheep TRB locus. *Mol. Immunol.* 45, 493–509. doi: 10.1016/j.molimm.2007.05.023

Antonacci, R., Mineccia, M., Lefranc, M. P., Ashmaoui, H. M. E., Lanave, C., Piccinni, B., et al. (2011). Expression and genomic analyses of Camelus dromedarius T cell receptor delta (TRD) genes reveal a variable domain repertoire enlargement due to CDR3 diversification and somatic mutation. *Mol. Immunol.* 48, 1384–1396. doi: 10.1016/j.molimm.2011.03.011

Antonacci, R., Giannico, F., Ciccarese, S., and Massari, S. (2014). Genomic characteristics of the T cell receptor (TRB) locus in the rabbit (Oryctolagus cuniculus) revealed by comparative and phylogenetic analyses. *Immunogenetics* 66, 255–266. doi: 10.1007/s00251-013-0754-1

Antonacci, R., Bellini, M., Pala, A., Mineccia, M., Hassanane, M. S., Ciccarese, S., et al. (2017a). The occurrence of three D-J-C clusters within the dromedary TRB locus highlights a shared evolution in Tylopoda, Ruminantia and Suina. *Dev. Comp. Immunol.* 76, 105–119. doi: 10.1016/j.dci.2017.05.021

Antonacci, R., Bellini, M., Castelli, V., Ciccarese, S., and Massari, S. (2017b). Data charactering the genomic structure of the T cell receptor (TRB) locus in Camelus dromedarius. *Data Brief* 14, 507–514. doi: 10.1016/j.dib.2017.08.002

Antonacci, R., Bellini, M., Ciccarese, S., and Massari, S. (2019). Comparative analysis of the TRB locus in the Camelus genus. *Front. Genet.* 10, 482. doi: 10.3389/fgene.2019.00482

Avila, F., Baily, M. P., Perelman, P., Das, P. J., Pontius, J., Chowdhary, R., et al. (2014). A comprehensive whole-genome integrated cytogenetic map for the alpaca (Lama pacos). *Cytogenet. Genome Res.* 144, 196–207. doi: 10.1159/000370329

Baudisch, B., Pfort, I., Sorge, E., and Conrad, U. (2018). Nanobody-Directed Specific Degradation of Proteins by the 26S-Proteasome in Plants. *Front. Plant Sci.* 9, 130. doi: 10.3389/fpls.2018.00130

Beghein, E., and Gettemans, J. (2017). Nanobody Technology: A versatile toolkit for microscopic imaging, protein-protein interaction analysis, and protein function exploration. *Front. Immunol.* 8, 771. doi: 10.3389/fimmu.2017.00771

Brooks, C. L., Rossotti, M. A., and Henry, K. A. (2018). Immunological functions and evolutionary emergence of heavy-chain antibodies. *Trends Immunol.* 39, 956–960. doi: 10.1016/j.it.2018.09.008

Carding, S. R., and Egan, P. J. (2002). Gammadelta T cells: functional plasticity and heterogeneity. *Nat. Rev. Immunol.* 2, 336–345. doi: 10.1038/nri797

Chen, H., Kshirsagar, S., Jensen, I., Lau, K., Covarrubias, R., Schluter, S. F., et al. (2009). Characterization of arrangement and expression of the T cell receptor gamma locus in the sandbar shark. *Proc. Natl. Acad. Sci. U. S. A.* 106, 8591–8596. doi: 10.1073/pnas.0811283106

Chen, H., Bernstein, H., Ranganathan, P., and Schluter, S. F. (2012). Somatic hypermutation of TCR γ V genes in the sandbar shark. *Dev. Comp. Immunol.* 37, 176–183. doi: 10.1016/j.dci.2011.08.018

Chothia, C., Novotný, J., Bruccoleri, R., and Karplus, M. (1985). Domain association in immunoglobulin molecules. The packing of variable domains. *J. Mol. Biol.* 186, 651–663. doi: 10.1016/0022-2836(85)90137-8

Ciccarese, S., Lanave, C., and Saccone, C. (1997). Evolution of T-cell receptors gamma and delta constant region and other T-cell related proteins in the human-rodent-artiodactyl triplet. *Genetics* 145, 409–419.

Ciccarese, S., Vaccarelli, G., Lefranc, M. P., Tasco, G., Consiglio, A., Casadio, R., et al. (2014). Characteristics of the somatic hypermutation in the Camelus dromedarius T cell receptor gamma (TRG) and delta (TRD) variable domains. *Dev. Comp. Immunol.* 46, 300–313. doi: 10.1016/j.dci.2014.05.001

Connelley, T., Aerts, J., Law, A., and Morrison, W. I. (2009). Genomic analysis reveals extensive gene duplication within the bovine TRB locus. *BMC Genomics* 10. doi: 10.1186/1471-2164-10-192

Conrad, M. L., Mawer, M. A., Lefranc, M. P., McKinnell, L., Whitehead, J., Davis, S. K., et al. (2007). The genomic sequence of the bovine T cell receptor gamma TRG loci and localization of the TRGC5 cassette. *Vet. Immunol. Immunopathol.* 115, 346–356. doi: 10.1016/j.vetimm.2006.10.019

Davies, J., and Riechmann, L. (1994). Camelising' human antibody fragments: NMR studies on VH domains. *FEBS Lett.* 339, 285–290. doi: 10.1016/0014-5793(94)80432-X

Decanniere, K., Desmyter, A., Lauwereys, M., Ghahroudi, M. A., Muyldermans, S., and Wyns, L. (1999). A single-domain antibody fragment in complex with RNase A: non-canonical loop structures and nanomolar affinity using two CDR loops. *Structure* 7, 361–370. doi: 10.1016/S0969-2126(99)80049-5

De Genst, E., Saerens, D., Muyldermans, S., and Conrath, K. (2006). Antibody repertoire development in camelids. *Dev. Comp. Immunol.* 30, 187–198. doi: 10.1016/j.dci.2005.06.010

Deschacht, N., De Groeve, K., Vincke, C., Raes, G., De Baetselier, P., and Muyldermans, S. (2010). A novel promiscuous class of camelid single-domain antibody contributes to the antigen-binding repertoire. *J. Immunol.* 184, 5696–5704. doi: 10.4049/jimmunol.0903722

Desmyter, A., Transue, T. R., Ghahroudi, M. A., Thi, M. H., Poortmans, F., Hamers, R., et al. (1996). Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. *Nat. Struct. Biol.* 3, 803–811. doi: 10.1038/nsb0996-803

Di Tommaso, S., Antonacci, R., Ciccarese, S., and Massari, S. (2010). Extensive analysis of D-J-C arrangements allows the identification of different mechanisms enhancing the diversity in sheep T cell receptor beta-chain repertoire. *BMC Genomics* 11. doi: 10.1186/1471-2164-11-3

Eguchi-Ogawa, T., Toki, D., and Uenishi, H. (2009). Genomic structure of the whole D-J-C clusters and the upstream region coding V segments of the TRB locus in pig. *Dev. Comp. Immunol.* 33, 1111–1119. doi: 10.1016/j.dci.2009.06.006

Ehrenmann, F., Kaas, Q., and Lefranc, M. P. (2010). IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.* 38, D301–D307. doi: 10.1093/nar/gkp946

Ehrenmann, F., Giudicelli, V., Duroux, P., and Lefranc, M. P. (2011). IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb. Protoc.* 6, 726–736. doi: 10.1101/pdb.prot5635

Fernandes, J. C. (2018). Therapeutic application of antibody fragments in autoimmune diseases: current state and prospects. *Drug Discov. Today* 23, 1996–2002. doi: 10.1016/j.drudis.2018.06.003

Fitak, R. R., Mohandesan, E., Corander, J., and Burger, P. A. (2016a). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.* 16, 314–324. doi: 10.1111/1755-0998.12443

Fitak, R., Mohandesan, E., Corander, J., Yadamsuren, A., Chuluunbat, B., Abdelhadi, O., et al. (2016b). Genomic Footprints of Selection Under Domestication in Old World Camelids. *Plant Animal Genomic Conf. XXIV*. San Diego.

Flajnik, M. F., Deschacht, N., and Muyldermans, S. (2011). A case of convergence: why did a simple alternative to canonical antibodies arise in sharks and camels? *PLoS Biol.* 9, e1001120. doi: 10.1371/journal.pbio.1001120

Govaert, J., Pellis, M., Deschacht, N., Vincke, C., Conrath, K., Muyldermans, S., et al. (2012). Dual beneficial effect of interloop disulfide bond for single domain antibody fragments. *J. Biol. Chem.* 287, 1970–1979. doi: 10.1074/jbc.M111.242818

Griffin, L. M., Snowden, J. R., Lawson, A. D., Wernery, U., Kinne, J., and Baker, T. S. (2014). Analysis of heavy and light chain sequences of conventional camelid antibodies from Camelus dromedarius and Camelus bactrianus species. *J. Immunol. Methods* 405, 35–46. doi: 10.1016/j.jim.2014.01.003

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hamers, C., Songa, E. B., et al. (1993). Naturally occurring antibodies devoid of light chains. *Nature* 363, 446–448. doi: 10.1038/363446a0

Harmsen, M. M., Ruuls, R. C., Nijman, I. J., Niewold, T. A., Frenken, L. G., and de Geus, B. (2000). Llama heavy-chain V regions consist of at least four distinct subfamilies revealing novel sequence features. *Mol. Immunol.* 37, 579–590. doi: 10.1016/S0161-5890(00)00081-X

Hassanzadeh-Ghassabeh, G., Devoogdt, N., De Pauw, P., Vincke, C., and Muyldermans, S. (2013). Nanobodies and their potential applications. *Nanomedicine* 8, 1013–1026. doi: 10.2217/nnm.13.86

Hein, W. R., and Dudler, L. (1993). Divergent evolution of T cell repertoires: extensive diversity and developmentally regulated expression of the sheep gamma delta T cell receptor. *EMBO J.* 12, 715–724. doi: 10.1002/j.1460-2075.1993.tb05705.x

Hein, W. R., and Dudler, L. (1997). TCR gamma delta cells are prominent in normal bovine skin and express a diverse repertoire of antigen receptors. *Immunology* 91, 58–64. doi: 10.1046/j.1365-2567.1997.00224.x

Helma, J., Cardoso, M. C., Muyldermans, S., and Leonhardt, H. (2015). Nanobodies and recombinant binders in cell biology. *J. Cell Biol.* 209, 633–644. doi: 10.1083/jcb.201409074

Henry, K. A., van Faassen, H., Harcus, D., Marcil, A., Hill, J. J., Muyldermans, S., et al. (2019). Llama peripheral B-cell populations producing conventional and heavy chain-only IgG subtypes are phenotypically indistinguishable but immunogenetically distinct. *Immunogenetics* 71, 307–320. doi: 10.1007/s00251-018-01102-9

Ishiguro, N., Aida, Y., Shinagawa, T., and Shinagawa, M. (1993). Molecular structures of cattle T-cell receptor gamma and delta chains predominantly expressed on peripheral blood lymphocytes. *Immunogenetics* 38, 437–443. doi: 10.1007/BF00184524

Janeway, C.A. Jr, Travers, P., Walport, M., et al. (2001). "Immunobiology: The Immune System in Health and Disease," in *The major histocompatibility complex and its functions*, (New York: Garland Science). Available from: https://www.ncbi.nlm.nih.gov/books/NBK27156/.

Jirimutu, Wang. Z., Ding, G., Chen, G., Sun, Y., Sun, Z., et al. (2012). Genome sequences of wild and domestic bactrian camels. *Nat. Commun.* 3, 1202. doi: 10.1038/ncomms2192

Jung, D., and Alt, F. W. (2004). Unraveling V(D)J recombination; insights into gene regulation. *Cell* 116, 299–311. doi: 10.1016/S0092-8674(04)00039-X

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kazen, A. R., and Adams, E. J. (2011). Evolution of the V, D, and J gene segments used in the primate gammadelta T-cell receptor reveals a dichotomy of conservation and diversity. *Proc. Natl. Acad. Sci. U. S. A.* 108, 332–340. doi: 10.1073/pnas.1105105108

Lauwereys, M., Arbabi Ghahroudi, M., Desmyter, A., Kinne, J., Hölzer, W., De Genst, E., et al (1998). Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *EMBO J.* 17, 3512–3520. doi: 10.1093/emboj/17.13.3512

Lefranc, M. P., and Rabbitts, T. H. (1989). The human T-cell receptor gamma (TRG) genes. *Trends Biochem. Sci.* 14, 214–218. doi: 10.1016/0968-0004(89)90029-7

Lefranc, M. P., and Lefranc, G. (2001a). *The Immunoglobulin Facts-Book*. Academic, New York 1–457.

Lefranc, M. P., and Lefranc, G. (2001b). *The T cell Receptor Facts-Book*. Academic, New York 1–398.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., et al. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V- like domains. *Dev. Comp. Immunol.* 27, 55–77. doi: 10.1016/S0145-305X(02)00039-3

Lefranc, M. P., Pommie´, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., et al. (2005). IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol.* 29, 185–203. doi: 10.1016/j.dci.2004.07.003

Lefranc, M. P. (2011a). IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb. Protoc.* 6, 633–642. doi: 10.1101/pdb.ip85

Lefranc, M. P. (2011b). IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb. Protoc.* 6, 643–651. doi: 10.1101/pdb.ip86

Lefranc, M. P. (2014a). Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. *Front. Immunol.* 5, 22. doi: 10.3389/fimmu.2014.00022

Lefranc, M. P. (2014b). "IMGT® immunoglobulin repertoire analysis and antibody humanization," in *Molecular Biology of B cells*, Eds. Alt, F. W., Honjo, T., Radbruch, A., and Reth, M. (London, UK: Academic Press, Elsevier Ltd), 481–514. doi: 10.1016/B978-0-12-397933-9.00026-6

Lefranc, M. P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., et al. (2015). IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* 43, D413–D422. doi: 10.1093/nar/gku1056

Lefranc, M. P., and Lefranc, G. (2019). "IMGT® and 30 years of Immunoinformatics Insight in antibody V and C domain structure and function," in *Antibodies*, vol. 8 . Eds. Jefferis, R., Strohl, W. R., and Kato, K., 29. doi: 10.3390/antib8020029

Li, X., Duan, X., Yang, K., Zhang, W., Zhang, C., Fu, L., et al (2016). Comparative analysis of immune repertoires between bactrian camel's conventional and heavy-chain antibodies. *PLoS One* 11, e0161801. doi: 10.1371/journal.pone.0161801

Li, Z., Woo, C. J., Iglesias-Ussel, M. D., Ronai, D., and Scharff, M. D. (2004). The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev.* 18, 1–11. doi: 10.1101/gad.1161904

Linguiti, G., Antonacci, R., Tasco, G., Grande, F., Casadio, R., Massari, S., et al. (2016). Genomic and expression analyses of Tursiops truncatus T cell receptor gamma (TRG) and alpha/delta (TRA/TRD) loci reveal a similar basic public γδ repertoire in dolphin and human. *BMC Genomics* 17, 634–651. doi: 10.1186/s12864-016-2841-9

Liu, M., and Schatz, D. G. (2009). Balancing AID and DNA repair during somatic hypermutation. *Trends Immunol.* 30, 173–181. doi: 10.1016/j.it.2009.01.007

Loris, R., Marianovsky, I., Lah, J., Laeremans, T., Engelberg-Kulka, H., Glaser, G., et al.. (2003). Crystal structure of the intrinsically flexible addiction antidote MazE. *J. Biol. Chem.* 278, 28252–28257. doi: 10.1074/jbc.M302336200

Massari, S., Lipsi, M. R., Vonghia, G., Antonacci, R., and Ciccarese, S. (1998). T-cell receptor TRG1 and TRG2 clusters map separately in two different regions of sheep chromosome 4. *Chromosome Res.* 6, 419–420. doi: 10.1023/A:1009245830804

Massari, S., Antonacci, R., Lanave, C., and Ciccarese, S. (2000). Genomic organization of sheep TRDJ segments and their expression in the delta chain repertoire in thymus. *Immunogenetics* 52, 1–8. doi: 10.1007/s002510000243

Massari, S., Bellahcene, F., Vaccarelli, G., Carelli, G., Mineccia, M., Lefranc, M. P., et al. (2009). The deduced structure of the T cell receptor gamma locus in Canis lupus familiaris. *Mol. Immunol.* 46, 2728–2736. doi: 10.1016/j.molimm.2009.05.008

Massari, S., Bellini, M., Ciccarese, S., and Antonacci, R. (2018). Overview of the germline and expressed repertoires of the TRB genes in Sus scrofa. *Front. Immunol.* 9, 2526. doi: 10.3389/fimmu.2018.02526

McWilliam, H. E., Eckle, S. B., Theodossis, A., Liu, L., Chen, Z., Wubben, J. M., et al. (2016). The intracellular pathway for the presentation of vitamin B-related antigens by the antigen-presenting molecule MR1. *Nat. Immunol.* 17, 531–537. doi: 10.1038/ni.3416

Miccoli, M. C., Lipsi, M. R., Massari, S., Lanave, C., and Ciccarese, S. (2001). Exon-intron organization of TRGC genes in sheep. *Immunogenetics* 53, 416–422. doi: 10.1007/s002510100340

Miccoli, M. C., Antonacci, R., Vaccarelli, G., Lanave, C., Massari, S., Cribiu, E. P., et al. (2003). Evolution of TRG clusters in cattle and sheep genomes as drawn from the structural analysis of the ovine TRG2 @ locus. *J. Mol. Evol.* 57, 52–62. doi: 10.1007/s00239-002-2451-9

Mineccia, M., Massari, S., Linguiti, G., Ceci, L., Ciccarese, S., and Antonacci, R. (2012). New insight into the genomic structure of dog T cell receptor beta (TRB) locus inferred from expression analysis. *Dev. Comp. Immunol.* 37, 279–293. doi: 10.1016/j.dci.2012.03.010

Mitchell, L. S., and Colwell, L. J. (2018a). Analysis of nanobody paratopes reveals greater diversity than classical antibodies. *Protein Eng. Des. Sel.* 31, 267–275. doi: 10.1093/protein/gzy017

Mitchell, L. S., and Colwell, L. J. (2018b). Comparative analysis of nanobody sequence and structure data. *Proteins* 86, 697–706. doi: 10.1002/prot.25497

Muyldermans, S., Atarhouch, T., Saldanha, J., Barbosa, J. A., and Hamers, R. (1994). Sequence and structure of VH domain from naturally occurring camel heavy chain immunoglobulins lacking light chains. *Protein Eng.* 7, 1129–1135. doi: 10.1093/protein/7.9.1129

Muyldermans, S., Baral, T. N., Retamozzo, V. C., De Baetselier, P., De Genst, E., Kinne, J., et al. (2009). Camelid immunoglobulins and nanobody technology. *Vet. Immunol. Immunopathol.* 128, 178–183. doi: 10.1016/j.vetimm.2008.10.299

Muyldermans, S., and Lauwereys, M. (1999). Unique single-domain antigen binding fragments derived from naturally occurring camel heavy-chain antibodies. *J. Mol. Recogn.* 12, 131–140. doi: 10.1002/(SICI)1099-1352(199903/04)12:2<131::AID-JMR454>3.0.CO;2-M

Muyldermans, S. (2013). Nanobodies: natural single-domain antibodies. *Annu. Rev. Biochem.* 82, 775–797. doi: 10.1146/annurev-biochem-063011-092449

Nguyen, V. K., Hamers, R., Wyns, L., and Muyldermans, S. (1999). Loss of splice consensus signal is responsible for the removal of the entire C(H)1 domain of the functional camel IGG2A heavy-chain antibodies. *Mol. Immunol.* 36, 515–524. doi: 10.1016/S0161-5890(99)00067-X

Nguyen, V. K., Hamers, R., Wyns, L., and Muyldermans, S. (2000). Camel heavy-chain antibodies: diverse germline V(H)H and specific mechanisms enlarge the antigen-binding repertoire. *EMBO J.* 19, 921–930. doi: 10.1093/emboj/19.5.921

Nguyen, V. K., Desmyter, A., and Muyldermans, S. (2001). Functional heavy-chain antibodies in Camelidae. *Adv. Immunol.* 79, 261–296. doi: 10.1016/S0065-2776(01)79006-2

Ott, J. A., Castro, C. D., Deiss, T. C., Ohta, Y., Flajnik, M. F., and Criscitiello, M. F. (2018). Somatic hypermutation of T cell receptor α chain contributes to selection in nurse shark thymus. *Elife* 17, 7. doi: 10.7554/eLife.28477

Pavlov, Y. I., Rogozin, I. B., Galkin, A. P., Aksenova, A. Y., Hanaoka, F., Rada, C., et al. (2002). Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase g during copying of a mouse immunoglobulin j light chain transgene. *Proc. Natl. Acad. Sci. U. S. A.* 99, 9954–9959. doi: 10.1073/pnas.152126799

Plasil, M., Mohandesan, E., Fitak, R. R., Musilova, P., Kubickova, S., Burger, P. A., et al. (2016). The major histocompatibility complex in Old World camelids and low polymorphism of its class II genes. *BMC Genomics* 17, 167. doi: 10.1186/s12864-016-2500-1

Plasil, M., Wijkmark, S., Elbers, J. P., Oppelt, J., Burger, P. A., and Horin, P. (2019). The major histocompatibility complex of Old World camelids: class I and class I-related genes. *HLA* 93, 203–215. doi: 10.1111/tan.13510

Rissiek, B., Koch-Nolte, F., and Magnus, T. (2014). Nanobodies as modulators of inflammation: potential applications for acute brain injury. *Front. Cell Neurosci.* 8, 344. doi: 10.3389/fncel.2014.00344

Rogozin, I. B., and Diaz, M. (2004). Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J. Immunol.* 172, 3382–3384. doi: 10.4049/jimmunol.172.6.3382

Saccodossi, N., De Simone, E. A., and Leoni, J. (2012). Structural analysis of effector functions related motifs, complement activation and hemagglutinating activities in Lama glama heavy chain antibodies. *Vet. Immunol. Immunopathol.* 145, 323–331. doi: 10.1016/j.vetimm.2011.12.001

Schanz, S., Castor, D., Fischer, F., and Jiricny, J. (2009). Interference of mismatch and base excision repair during the processing of adjacent U/G mispairs may play a key role in somatic hypermutation. *Proc. Natl. Acad. Sci. U. S. A.* 106, 5593–5598. doi: 10.1073/pnas.0901726106

Schumacher, D., Helma, J., Schneider, A. F. L., Leonhardt, H., and Hackenberger, C. P. R. (2018). Nanobodies: chemical functionalization strategies and intracellular applications. *Angew. Chem. Int. Ed. Engl.* 57, 2314–2333. doi: 10.1002/anie.201708459

Shafi, S., Vantourout, P., Wallace, G., Antoun, A., Vaughan, R., Stanford, M., et al.. (2011). An NKG2D-mediated human lymphoid stress surveillance response with high interindividual variation. *Sci. Transl. Med.* 3, 113ra124. doi: 10.1126/scitranslmed.3002922

Sironi, M., Cagliani, R., Forni, D., and Clerici, M. (2015). Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nat. Rev. Gen.* 16, 224. doi: 10.1038/nrg3905

Steeland, S., Vandenbroucke, R. E., and Libert, C. (2016). Nanobodies as therapeutics: big opportunities for small antibodies. *Drug Discov. Today* 21, 1076–1113. doi: 10.1016/j.drudis.2016.04.003

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi: 10.1093/oxfordjournals.molbev.a040023

Tillib, S. V., Vyatchanin, A. S., and Muyldermans, S. (2014). Molecular analysis of heavy chain-only antibodies of Camelus bactrianus. *Biochemistry* 79, 1382–1390. doi: 10.1134/S000629791412013X

Trowsdale, J., and Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. *Ann. Rev. Gen. Hum. Gen.* 14, 301–323. doi: 10.1146/annurev-genom-091212-153455

Vaccarelli, G., Miccoli, M. C., Lanave, C., Massari, S., Cribiu, E. P., and Ciccarese, S. (2005). Genomic organization of the sheep TRG1{{{at}}} locus and comparative analyses of Bovidae and human variable genes. *Gene* 357, 103–114. doi: 10.1016/j.gene.2005.05.033

Vaccarelli, G., Miccoli, M. C., Antonacci, R., Pesole, G., and Ciccarese, S. (2008). Genomic organization and recombinational unit duplication-driven evolution of ovine and bovine T cell receptor gamma loci. *BMC Genomics* 9, 81. doi: 10.1186/1471-2164-9-81

Vaccarelli, G., Antonacci, R., Tasco, G., Yang, F. T., Giordano, L., El Ashmaoui, H. M., et al. (2012). Generation of diversity by somatic mutation in the Camelus dromedarius T-cell receptor gamma variable domains. *Eur. J. Immunol.* 42, 3416–3428. doi: 10.1002/eji.201142176

Vu, K. B., Ghahroudi, M. A., Wyns, L., and Muyldermans, S. (1997). Comparison of llama VH sequences from conventional and heavy chain antibodies. *Mol. Immunol.* 34, 1121–1131. doi: 10.1016/S0161-5890(97)00146-6

Wilson, T. M., Vaisman, A., Martomo, S. A., Sullivan, P., Lan, L., Hanaoka, F., et al. (2005). MSH2-MSH6 stimulates DNA polymerase eta, suggesting a role for A: T mutations in antibody genes. *J. Exp. Med.* 201, 637–645. doi: 10.1084/jem.20042066

Woolven, B. P., Frenken, L. G., van der Logt, P., and Nicholls, P. J. (1999). The structure of the llama heavy chain constant genes reveals a mechanism for heavy-chain antibody formation. *Immunogenetics* 50, 98–101. doi: 10.1007/s002510050694

Wu, H., Guang, X., Al-Fageeh, M. B., Cao, J., Pan, S., Zhou, H., et al.. (2014). Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.* 5, 5188. doi: 10.1038/ncomms6188

Xiao, G., Wang, X., Sheng, J., Lu, S., Yu, X., and Wu, J. D. (2015). Soluble NKG2D ligand promotes MDSC expansion and skews macrophage to the alternatively activated phenotype. *J. Hematol. Oncol.* 8, 13. doi: 10.1186/s13045-015-0110-z

Xu, B., Pizarro, J. C., Holmes, M. A., McBeth, C., Groh, V., Spies, T., et al. (2011). Crystal structure of a gammadelta T-cell receptor specific for the human MHC χλασσ I homolog MICA. *Proc. Natl. Acad. Sci. U. S. A.* 108, 2414–2419. doi: 10.1073/pnas.1015433108

Yang, Y. G., Ohta, S., Yamada, S., Shimizu, M., and Takagaki, Y. (1995). Diversity of T cell receptor delta-chain cDNA in the thymus of a one-month-old pig. *J. Immunol.* 155, 1981–1993.

Zhao, Y., Gregory, M. T., Biertümpfel, C., Hua, Y.-J., Hanaoka, F., and Yangb, W. (2013). Mechanism of somatic hypermutation at the WA motif by human DNA polimerase g. *Proc. Natl. Acad. Sci. U. S. A.* 110, 8146–8151. doi: 10.1073/pnas.1303126110

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Ciccarese, Burger, Ciani, Castelli, Linguiti, Plasil, Massari, Horin and Antonacci. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership