GENETIC VARIABILITY IN CONSERVATION AND SELECTION PROGRAMS IN THE POST-GENOMICS ERA

EDITED BY: Maria Saura, Francesca Bertolini and Christos Palaiokostas
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-249-4
DOI 10.3389/978-2-88966-249-4

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

GENETIC VARIABILITY IN CONSERVATION AND SELECTION PROGRAMS IN THE POST-GENOMICS ERA

Topic Editors:
Maria Saura, Instituto Nacional de Investigación y Tecnología Agroalimentaria (INIA), Spain
Francesca Bertolini, Technical University of Denmark, Denmark
Christos Palaiokostas, Swedish University of Agricultural Sciences, Sweden

Citation: Saura, M., Bertolini, F., Palaiokostas, C., eds. (2020). Genetic Variability in Conservation and Selection Programs in the Post-Genomics Era. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-249-4

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2019
doi: 10.3389/fgene.2019.00665

Comparative Genomic Analysis of Three Salmonid Species Identifies Functional Candidate Genes Involved in Resistance to the Intracellular Bacterium *Piscirickettsia salmonis*

José M. Yáñez[1,2], Grazyella M. Yoshida[1], Ángel Parra[1,3,4,5], Katharina Correa[6], Agustín Barría[1,7], Liane N. Bassini[8], Kris A. Christensen[9], Maria E. López[1,10], Roberto Carvalheiro[11,12], Jean P. Lhorente[6] and Rodrigo Pulgar[3]**

[1] Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile, Santiago, Chile, [2] Núcleo Milenio INVASAL, Concepción, Chile, [3] Instituto de Nutrición y Tecnología de los Alimentos, Universidad de Chile, Santiago, Chile, [4] Doctorado en Acuicultura. Programa Cooperativo Universidad de Chile, Universidad Católica del Norte, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile, [5] Facultad de Ciencias del Mar, Universidad Católica del Norte, Coquimbo, Chile, [6] Benchmark Genetics Chile, Puerto Montt, Chile, [7] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh Easter Bush, Midlothian, United Kingdom, [8] Escuela de Medicina Veterinaria, Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago, Chile, [9] Fisheries and Oceans Canada, West Vancouver, BC, Canada, [10] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden, [11] School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal, Brazil, [12] National Council for Scientific and Technological Development (CNPq), Brasília, Brazil

OPEN ACCESS

Edited by:
Christos Palaiokostas, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Alastair Hamilton, Hendrix Genetics BV, Netherlands
Marieke Verleih, Leibniz Institute for Farm Animal Biology, Germany

***Correspondence:**
José M. Yáñez
jmayanez@uchile.cl
Rodrigo Pulgar
rpulgar@inta.uchile.cl

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *24 March 2019*
Accepted: *25 June 2019*
Published: *05 August 2019*

Citation:
Yáñez JM, Yoshida GM, Parra Á, Correa K, Barría A, Bassini LN, Christensen KA, López ME, Carvalheiro R, Lhorente JP and Pulgar R (2019) Comparative Genomic Analysis of Three Salmonid Species Identifies Functional Candidate Genes Involved in Resistance to the Intracellular Bacterium Piscirickettsia salmonis. Front. Genet. 10:665. doi: 10.3389/fgene.2019.00665

Piscirickettsia salmonis is the etiologic agent of salmon rickettsial syndrome (SRS) and is responsible for considerable economic losses in salmon aquaculture. The bacterium affects coho salmon (CS; *Oncorhynchus kisutch*), Atlantic salmon (AS; *Salmo salar*), and rainbow trout (RT; *Oncorhynchus mykiss*) in several countries, including Norway, Canada, Scotland, Ireland, and Chile. We used Bayesian genome-wide association study analyses to investigate the genetic architecture of resistance to *P. salmonis* in farmed populations of these species. Resistance to SRS was defined as the number of days to death and as binary survival (BS). A total of 828 CS, 2130 RT, and 2601 AS individuals were phenotyped and then genotyped using double-digest restriction site-associated DNA sequencing and 57K and 50K Affymetrix® Axiom® single nucleotide polymorphism (SNP) panels, respectively. Both traits of SRS resistance in CS and RT appeared to be under oligogenic control. In AS, there was evidence of polygenic control of SRS resistance. To identify candidate genes associated with resistance, we applied a comparative genomics approach in which we systematically explored the complete set of genes adjacent to SNPs, which explained more than 1% of the genetic variance of resistance in each salmonid species (533 genes in total). Thus, genes were classified based on the following criteria: i) shared function of their protein domains among species, ii) shared orthology among species, iii) proximity to the SNP explaining the highest proportion of the genetic variance, and iv) presence in more than one genomic region explaining more than 1% of the genetic variance within species. Our results allowed us to identify 120 candidate genes belonging to at least one of the

four criteria described above. Of these, 21 of them were part of at least two of the criteria defined above and are suggested to be strong functional candidates influencing *P. salmonis* resistance. These genes are related to diverse biological processes, such as kinase activity, GTP hydrolysis, helicase activity, lipid metabolism, cytoskeletal dynamics, inflammation, and innate immune response, which seem essential in the host response against *P. salmonis* infection. These results provide fundamental knowledge on the potential functional genes underpinning resistance against *P. salmonis* in three salmonid species.

Keywords: coho salmon, rainbow trout, Atlantic salmon, *Piscirickettsia salmonis*, genome-wide association study, comparative genomics, piscirickettsiosis

INTRODUCTION

Infectious diseases are responsible for large economic losses in salmon farming. *Piscirickettsia salmonis*, the causal agent of salmon rickettsial syndrome (SRS), affects several salmon species and is considered one of the major pathogens affecting the salmon farming industry (Rozas and Enríquez, 2014). *P. salmonis* was identified in 1989 from farmed coho salmon (CS; *Oncorhynchus kisutch*) sampled in Chile (Cvitanich et al., 1991). Since then, *P. salmonis* has been confirmed as the causative agent for clinical and chronic SRS in CS, Atlantic salmon (AS; *Salmo salar*), and rainbow trout (RT; *Oncorhynchus mykiss*) in several countries, including Norway, Canada, Scotland, Ireland, and Chile (Fryer and Hedrick, 2003). Current control protocols and treatments are based on antibiotics and vaccines. The effectiveness of both strategies in field conditions is not optimal (Rozas and Enríquez, 2014). From the total mortalities ascribed to infectious diseases in Chile, SRS is responsible for 18.3%, 92.6%, and 67.9% in CS, RT, and AS, respectively (Sernapesca, 2018). These mortality rates, together with other factors such as antibiotic treatments and vaccinations, have generated economic losses up to USD $450 million per year (Camusetti et al., 2015).

A feasible and sustainable alternative to prevent disease outbreaks is genetic selection for disease resistance (Bishop and Woolliams, 2014). The estimated levels of heritability for resistance to *P. salmonis* in CS, AS, and RT range from 0.11 to 0.41 (Correa et al., 2015; Yáñez et al., 2016; Bangera et al., 2017; Barría et al., 2018a; Yoshida et al., 2018a; Bassini et al., 2019), demonstrating the feasibility of improving *P. salmonis* resistance through artificial selection in farmed salmon species.

Currently, the advancement of molecular technologies has allowed the generation of dense marker panels for salmonid species (Houston et al., 2014; Palti et al., 2015; Yañez et al., 2016; Macqueen et al., 2017). The use of genotypes from dense panels of single nucleotide polymorphism (SNP) markers, together with phenotypes for the traits of interest, assessed in a large number of individuals could provide opportunities to discover the genetic architecture of complex traits. When genetic markers are linked to a major effect of quantitative trait loci (QTL), marker-assisted selection (MAS) could then be implemented into breeding programs. For instance, a QTL explaining ~80% of the genetic variance for resistance to infectious pancreatic necrosis virus (IPNV) has been identified in Scottish and Norwegian AS farmed populations (Houston et al., 2008; Moen et al., 2009). To date, the number of IPNV outbreaks has been significantly reduced in Norwegian AS populations because of MAS for IPNV resistance (Hjeltnes et al., 2018). Interestingly, Moen et al. (2015) mapped the QTL to a region containing an epithelial cadherin (*cdh1*) gene encoding a protein that binds to IPNV, indicating that the protein is part of the machinery used by the virus for host internalization.

P. salmonis resistance has been suggested to be polygenic, with many loci explaining a small amount of the total genetic variance (Correa et al., 2015; Barría et al., 2018a), suggesting that the implementation of genomic selection (GS) is the most appropriate strategy to accelerate the genetic progress for this trait. Methods that can model all available SNPs simultaneously, including Bayesian regression methods (Fernando and Garrick, 2013), appear to be better for estimating marker effects than conventional methods of modeling each SNP individually and therefore are becoming increasingly more popular for genome-wide association study (GWAS; Goddard et al., 2009).

Due to the fact that *P. salmonis* affects farmed populations of three phylogenetically related salmonid species, including CS, AS, and RT, generating mortalities in a similar manner and that genetic variation for *P. salmonis* resistance has been already reported, we believe that exploring the genetic architecture of this trait simultaneously in the three species can provide further insights into the biology of the differential response against this intracellular bacteria among individuals. Thus, a comparative genomics approach aiming at evaluating and comparing genomic regions involved in *P. salmonis* resistance in CS, AS, and RT would help in narrowing down the list of potential candidate genes associated with the trait for further functional validation in salmonid species.

The aims of this study were i) to dissect the genetic architecture of resistance to *P. salmonis* in CS, AS, and RT using SNP and phenotype data modeled together using Bayesian GWAS approach, ii) to identify genomic regions involved in *P. salmonis* resistance among the three salmonid species, and iii) to identify candidate genes associated with *P. salmonis* resistance through a comparative genomics analysis.

MATERIALS AND METHODS

Challenge Tests

A total of 2,606, 2,601, and 2,416 fish belonging to 107, 118, and 105 full-sib families from CS, AS, and RT, respectively, were independently challenged with an isolate of *P. salmonis* (strain

LF-89; Mandakovic et al., 2016) as described in Barría et al. (2018a), Bassini et al. (2019), and Yáñez et al. (2013), Yáñez et al. (2014), Yáñez et al. (2016). Before the beginning of each experimental challenge, quantitative polymerase chain reaction (qPCR) was performed in a sub-sample of each population to confirm the absence of *Flavobacterium* spp., infectious salmon anemia virus, and IPNV. Subsequently, fish were intraperitoneally (IP) injected with 0.2 ml of an LD_{50} inoculum of *P. salmonis*. Although an IP challenge is not a natural form of infection, it is an effective method for presenting a naïve animal with a known and controlled amount of bacteria, making sure that the bacterial load and the time of infection are the same in every fish (Pulgar et al., 2015). After IP injection, infected fish were equally distributed by family into three different test tanks. Each challenge was maintained until mortalities returned to baseline levels. At the end of the challenges, all surviving fish were anesthetized and euthanized. A sample of caudal fin was taken from each survivor and dead fish from each of the experimental challenges for DNA extraction. Body weight was measured at the beginning of the challenge and at the time of death for each individual. The presence of *P. salmonis* was confirmed in a random sample of dead fish through qPCR and necropsy. Each experimental challenge was performed at Aquainnovo's Research Station, Xth Region, Chile.

Genotyping

A total of 828 CS, 2130 RT, and 2601 AS were genotyped using double-digest restriction site-associated DNA (ddRAD) and 57K and 50K Affymetrix® Axiom® SNP panels, respectively. Total DNA was extracted using commercial kits following the manufacturer's protocols. For CS, we used the Wizard SV Genomic DNA purification System (Promega), whereas DNeasy Blood & Tissue (Qiagen) was used for RT and AS.

For CS, 10 ddRAD libraries were prepared following the protocol proposed by Peterson et al. (2012) and sequenced on an Illumina Hiseq2500 (150 bp single-end). Raw sequences were analyzed using STACKS version 1.41 (Catchen et al., 2011; Catchen et al., 2013). rad-tags that passed the *process_radtags* quality control (QC) were aligned to the CS reference genome (GCF_002021735.1). Loci were built with *pstacks* setting a minimum depth coverage of three. After catalog construction, rad-tags were matched using *sstacks* followed by *populations* using default parameters. QC included the removal of SNPs below the following thresholds: Hardy-Weinberg equilibrium (HWE) $P < 1 \times 10^{-6}$, minor allele frequency (MAF) < 0.05, and genotyping call rate < 0.80. Individuals with a call rate below 0.70 were removed from the subsequent analysis. For a detailed protocol of library construction and SNP identification, see Barría et al. (2018a).

RT individuals were genotyped using the commercial 57K Affymetrix® Axiom® SNP array developed by the National Center of Cool and Cold Water Aquaculture at the U.S. Department of Agriculture (Palti et al., 2015). SNPs were filtered with the following QC parameters: HWE $P < 1 \times 10^{-6}$, MAF < 0.05, and SNP call rate < 0.95. Individuals with call rates lower than 0.95 were also removed.

The 50K Affymetrix® Axiom® SNP array used to genotype AS was developed by Universidad de Chile and Aquainnovo (Correa et al., 2015; Yañez et al., 2016). These markers were selected from a 200K array, as described in detail by Correa et al. (2015). Genotypes were subjected to QC using the following criteria: HWE $P < 1 \times 10^{-6}$, MAF < 0.05, SNP, and samples were discarded when the genotype rate was < 0.95.

GWAS analysis

Resistance to SRS was defined as both the number of days to death (DD) after experimental challenge and the binary survival (BS; 0 for surviving individuals at the end of the experimental challenge and 1 for deceased fish). GWAS analyses were performed using the Bayes C method that assumes distributed mixture distribution for marker effects. All model parameters are defined in the following equation:

$$y = Xb + Zu + \sum_{i=1}^{n} g_i a_i \delta_i + e \qquad \text{(EQ1)}$$

where y is the vector of phenotypic records (DD or BS); X and Z are the incidence matrix of fixed effects and polygenic effect, respectively; b is the vector of fixed effects (tank and body weight); u is the random vector of polygenic effects of all individuals in the pedigree; g_i is the vector of the genotypes for the *i*th SNP for each animal; a_i is the random allele substitution effect of the *i*th SNP; δ_i is an indicator variable (0, 1) sampled from a binomial distribution with parameters determined such that π value of 0.99; and *e* is a vector of residual effects.

The prior assumption is that SNP effects have independent and identical mixture distributions, where each SNP has a point mass at zero (with probability π) and a univariate Gaussian distribution (with probability 1 − π) with a mean equal to zero and variance equal to σ_a^2 having in turn a scaled inverse χ^2 prior, with $\nu_a = 4$ and $\nu_e = 10$ degrees of freedom and scale parameter, respectively (Fernando and Garrick, 2013). These hyperparameter values were chosen based on previous studies (Peters et al., 2012; Santana et al., 2016; Wolc et al., 2016; Yoshida et al., 2017; Yoshida et al., 2018a).

The analyses were performed using the GS3 software (Legarra et al., 2013). A total of 200,000 iterations in Gibbs sampling were used, with a burn-in period of 20,000 cycles, and the results were saved every 50 cycles. Convergence was assessed by visual inspection of trace plots of the posterior density of genetic and residual variances.

The proportion of the genetic variance explained (GEV) by each significant SNP was calculated as

$$Vg_i = \left(\frac{2 p_i q_i a_i^2}{\sigma_u^2} \right) \qquad \text{(EQ2)}$$

where p_i and q_i are the allele frequencies for the *i*th SNP, a_i is the estimated additive effect of the *i*th SNP on the phenotype, and σ_u^2 is the estimate of the polygenic variance (Lee et al., 2013).

The association between the SNPs and the phenotypes was assessed using the proportion of the GEV by each marker. To be inclusive regarding the genomic regions to be compared across the three species, we selected each of the regions explaining at least 1% of the genetic variance for the trait in each species.

The heritability values were calculated as

$$h^2 = \frac{V'_A}{V'_A + \sigma^2_e} \quad \text{(EQ3)}$$

where V'_A is the total additive genetic variance estimated as the sum of additive marker $\left(2\sigma^2_a \pi \sum p_i q_i\right)$ and the polygenic pedigree based $\left(\sigma^2_g\right)$ additive genetic variance.

Comparative Genomic Analysis

Initially, sequence homologies between chromosomes containing regions with SNPs explaining more than 1% of the genetic variance were compared. Synteny among these chromosomes was identified using Symap (Soderlund et al., 2011). The relationship between the chromosomes from CS, RT, and AS and the association between SNPs and resistance to *P. salmonis* (Manhattan plot) was plotted using Circos (Krzywinski et al., 2009).

To identify caandidate genes asstociated with *P. salmonis* resistance, we used a comparative genomic analysis among CS, RT, and AS. For this, we mapped the location of each SNP that explained 1% or more of the genetic variance for the trait on the reference genome (NCBI_RefSeq) of each species: CS (GCF_002021735.1), RT (GCF_002163495.1; Pearse et al., 2018), and AS (GCF_000233375.1; Lien et al., 2016). Subsequently, we retrieved the sequences of all the genes (and their protein products) adjacent to each SNP within a window of 1 Mb (500 kb downstream and 500 kb upstream to the associated SNP). We then used this information to apply the following criteria to classify and prioritize functional candidate genes by comparing the genomic regions involved in *P. salmonis* defined as DD and BS within and among the three species:

i) The complete set of genes was identified and classified into homologous superfamilies based on InterPro (Mitchell et al., 2019) protein domain signatures using Blast2GO software version 5.2.5 (Götz et al., 2008; referred to as Group A);
ii) Orthologous and paralogous genes among species were identified using the ProteinOrtho tool (Lechner et al., 2011). Multidirectional alignments were performed using the full-length sequences among complete sets of proteins encoded in each of the three species to obtain orthologous groups, with a 35% threshold for identity and similarity (Group B);
iii) The complete set of genes within 1 Mb windows adjacent to SNPs explaining the highest proportion of the genetic variation for each trait (leader SNP) was recovered and classified as high priority genes (Group C); and
iv) The complete set of genes located at the intersection of more than 1 Mb windows within a species was also identified and considered as high priority genes (Group D).

RESULTS

Challenge Test and Genetic Parameters

There was considerable phenotypic variation for *P. salmonis* resistance across fish species (**Figure 1**). The average cumulative mortality for different families ranged from 5% to 81%, 8% to 100%, and 8.3% to 73.7% for CS, RT, and AS, respectively. This

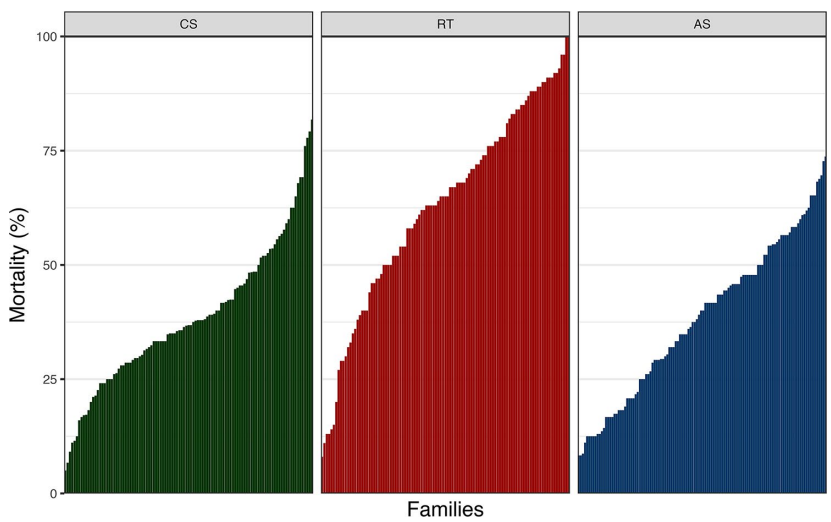


FIGURE 1 | Cumulative mortality by family after *P. salmonis* experimental infection of CS, RT, and AS. For CS, RT, and AS, a total of 107, 105, and 118 full-sib families were experimentally challenged.

TABLE 1 | Estimates of total additive genetic variance $\left(V_a'\right)$, residual variance $\left(\sigma_e^2\right)$, heritability (h^2), and standard deviation (SD) for resistance against *P. salmonis* in three salmonid species.

Species	DD				Binary survival			
	V_a'	σ_e^2	h^2	SD	V_a'	σ_e^2	h^2	SD
CS	28.91	60.70	0.32	0.07	7.53	1.00	0.88	0.03
RT	30.42	32.71	0.48	0.04	1.87	1.00	0.64	0.05
AS	16.52	53.17	0.24	0.04	0.47	1.00	0.32	0.05

result suggests that the phenotypic variation for this trait could be related to the genetic background on each species. Estimated heritabilities for *P. salmonis* resistance were significant for the three species, indicating the feasibility to improve the trait by means of artificial selection (**Table 1**). The genomic heritability values for DD were 0.32 for CS, 0.48 for RT, and 0.24 for AS. When resistance was defined as BS, genomic heritability estimates increased to 0.88, 0.64, and 0.32 for CS, RT, and AS, respectively, representing moderate to high levels of genetic variation for *P. salmonis* resistance.

GWAS Analysis

A total of 580 CS (9,389 SNPs), 1,929 RT (24,916 SNPs), and 2,383 AS (42,624 SNPs) were retained after QC. For CS and RT, we found relatively few SNPs explaining a moderate to high percentage of genetic variance for *P. salmonis* resistance. In contrast, for AS, a large number of SNPs with small effect were found and the percentage of GEV by a single marker was not higher than 5% (**Figure 2**; **Supplementary Figure 1**). Although there were multiple shared syntenic regions with associated SNPs (4 for DD and 5 for BS) in two species, there were no shared syntenic regions where all three species had common associated SNPs (**Figure 2**). **Figure 3** (and **Supplementary Figure 2**) highlights the different genetic architecture for resistance to *P. salmonis* among the three salmonid species studied. For CS, the top 200 SNPs explained about 70% and 90% of genetic variance for DD and BS, respectively, and just a marker located in chromosome 29 represented more than 50% of total genetic variance for BS. For RT, the top 200 SNPs explained 90% and 80% for DD and BS, respectively, whereas, in AS, they explained slightly more than 30% for both traits. These results suggested that CS and RT both appear to have oligogenic control with few markers having large effect loci, whereas the small effect of loci suggested the polygenic nature for resistance to *P. salmonis* in AS.

Comparative Genomic Analysis

We mapped the location of each SNP that explained 1% or more of the genetic variance for both DD and BS to the reference genome of CS, RT, and AS and searched for genes within 1 Mb windows flanking each SNP. This search allowed us to identify 533 unique genes that encoded 957 proteins. The complete list of genes and proteins can be found in **Supplementary Table S1: Sheets 1 to 6.**

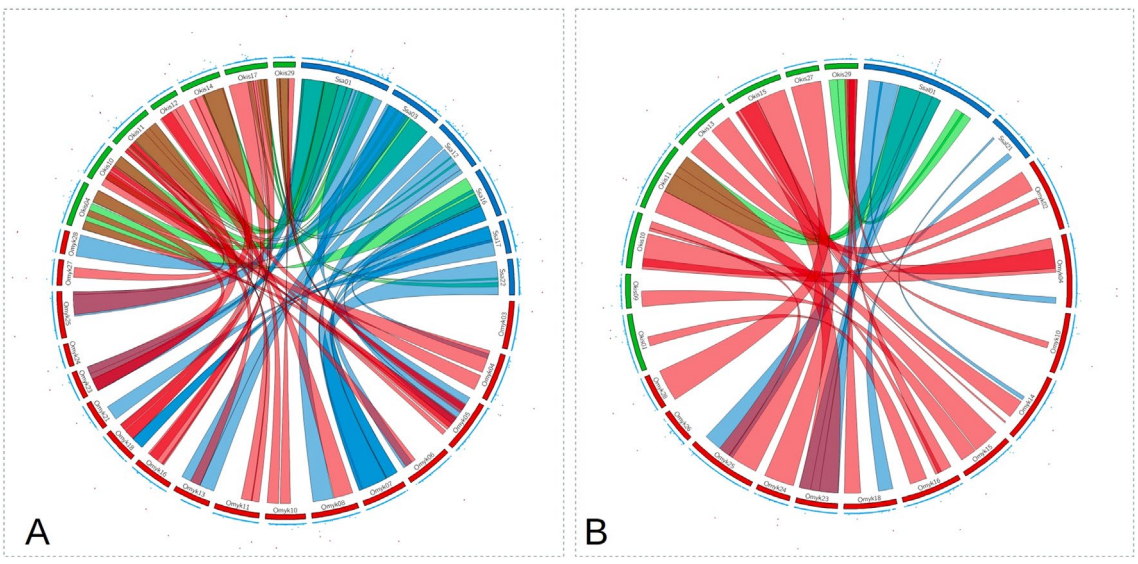


FIGURE 2 | Circos plot for *P. salmonis* resistance as DD **(A)** and as BS **(B)**. The inner ribbons mark syntenic regions among CS (green), RT (red), and AS (blue). Manhattan plots are shown on the outer ring, with significant associations plotted in red (values ≥ 1).

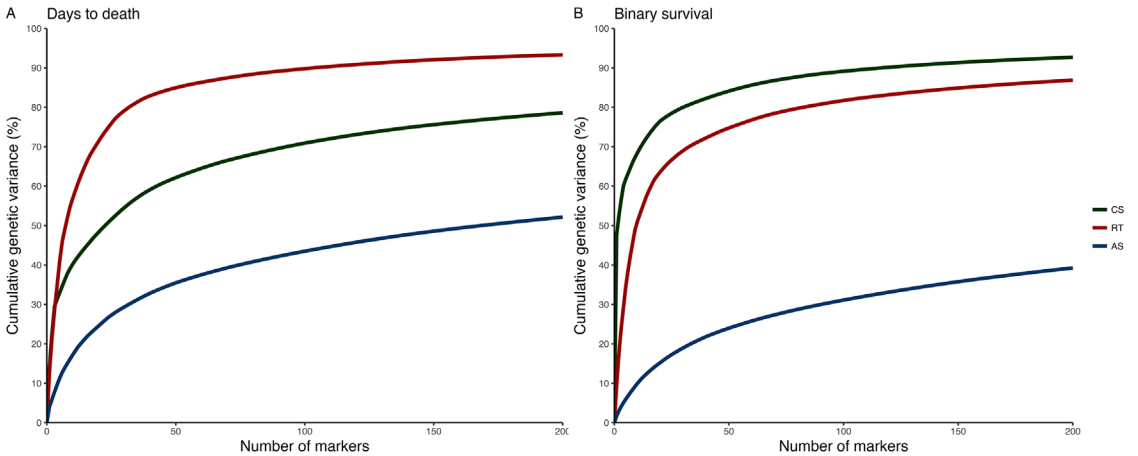


FIGURE 3 | Cumulative percentage of the GEV by the top 200 markers from Bayesian GWAS for resistance to *P. salmonis* measured as DD **(A)** and BS **(B)** in CS, RT, and AS.

To prioritize functional candidate genes, we annotated and classified the complete set of encoded proteins in homologous superfamilies for each trait and species based on InterPro protein domain signatures. We identified 194 and 129 homologous superfamilies for DD and BS, respectively, 103 of which were shared between traits (**Supplementary Table S1: homologous superfamilies**). The homologous superfamilies and the number of proteins present in at least two salmonid species are shown in **Figure 4**. Remarkably, about 30% of the proteins from genes present in regions associated with DD belong to five homologous superfamilies [*P-loop containing nucleoside triphosphate hydrolase* (also known as P-loop_NTPase), *immunoglobulin-like fold*, *zinc finger C2H2 superfamily*, *zinc finger RING/FYVE/PHD-type*, *and protein kinase-like domain superfamily*]. A total of 30% of proteins from genes present in regions associated with BS belong to only three homologous superfamilies (*P-loop_NTPase*, *immunoglobulin-like fold*, and *immunoglobulin-like domain superfamily*). Interestingly, the P-loop_NTPase superfamily contained the largest group of proteins for both traits, and at least one representative protein from each salmonid species belonged

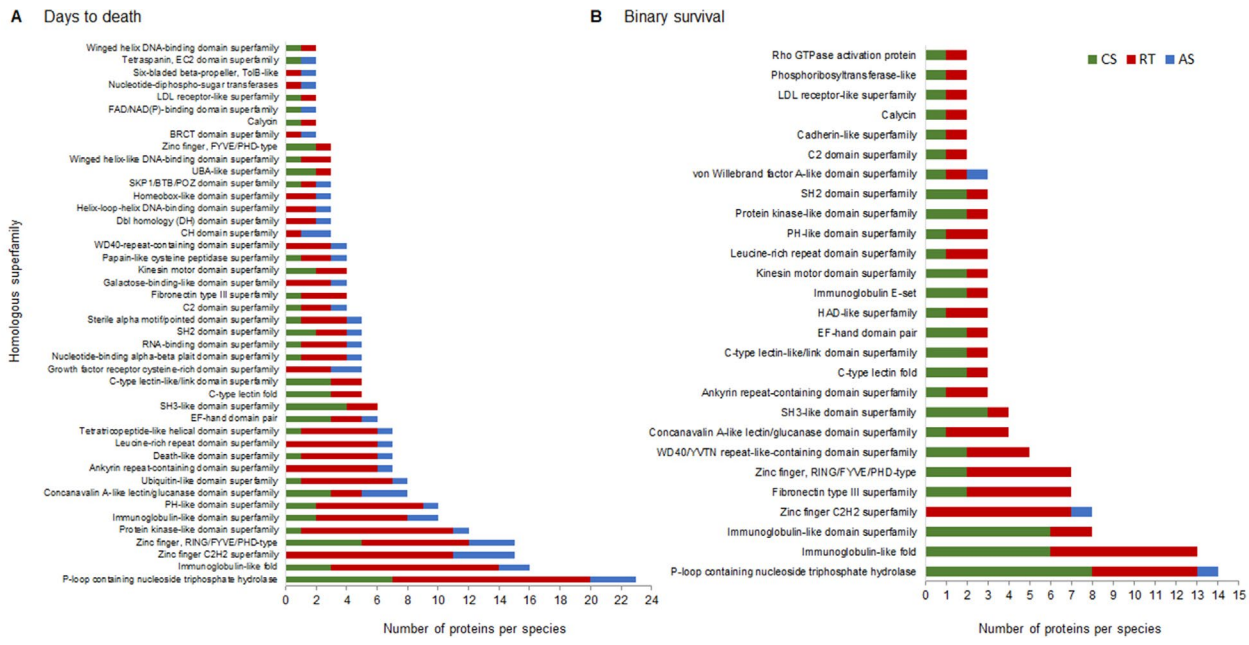


FIGURE 4 | Homologous superfamilies (InterPro) adjacent to the complete set of SNPs that explain more than 1% of the genetic variance of resistance to SRS measured as DD **(A)** and BS **(B)**. Bars represent the abundance of genes in each homologous superfamily present in at least two salmonid species. CS, RT, and AS.

to this superfamily. Thirty-one of the proteins identified in this study are part of this superfamily, including some GTPases, kinesin and myosin proteins, and ATP-dependent RNA helicases [**Supplementary Table S1, sheet: P-loop NTPases** (**Group_A**)].

To complement these analyses, we looked for orthologous proteins through multi-directional alignments using full-length sequences of the complete set of proteins for each species (Group B). Only five groups of orthologous genes were identified in at least two species, highlighting three non-receptor tyrosine-protein kinases (nr-TPK) with representative genes in the three species for DD and two species for BS. In addition, for DD, two ATP-dependent RNA helicases (DDX) and two Ras-related proteins (RAB) were identified in CS and RT, whereas two FYVE, RhoGEF/PH domain-containing proteins (FGD) were identified in RT and AS. For BS, two fatty acid-binding proteins (L-FABP) and two ankyrin repeat domain-containing proteins were identified in CS and RT [**Supplementary Table S1, sheet: Orthologous genes** (**Group_B**)]**.** The proteins nr-TPK, DDX, and L-FABP are also encoded by genes adjacent to SNPs that explained the highest proportion for the genetic variance (leader SNP) for both trait definitions (Group C).

Group C contained other genes (n=42) that encoded proteins such as myosin-IIIb (MYO3B), ATP-dependent RNA helicase (TDRD9), kinesin protein (KIF15), and kinesin protein (KIF2C) that are also included into the P-loop_NTPase superfamily as well as members of the orthologous groups such as FABP. Other genes encoding proteins classically associated with immune response such as tripartite motif-containing protein 35 (TRIM35) and lysozyme C II (LYZ2) are also part of this group. A complete list of these genes and proteins is in **Supplementary Table S1, sheet: Adjacent to leader SNP** (**Group_C**).

Group D was composed of genes (n=58) located adjacent to more than one SNP simultaneously (within overlapped windows). Among them, we identified GTPase IMAP family member 4 (GIMAP4), GTPase IMAP family member 8 (GIMAP8), NLR family CARD domain-containing protein 3 (NLRC3), ADP-ribosylation factor protein 5B (ARL5B), voltage-dependent L-type calcium channel subunit beta-2 (CACNB2), and heparan sulfate glucosamine 3-O-sulfotransferase 3A1 (HS3ST3A1), all of which are also P-loop_NTPases. In addition, we identified histidine triad nucleotide-binding protein 1 (HINT1), which is also adjacent to the leader SNP for DD in AS, and other genes associated with immune response such as collectin-12 (COL12), macrophage mannose receptor 1 (MRC1), and tapasin-related protein (TAPBPR). A complete list of these genes and proteins can be found in **Supplementary Table S1, sheet: Genes overlapped windows** (**Group_D**)**.** Additionally, the gene that codes for NACHT, LRR, and PYD domains-containing protein 12 (NLRP12) was found in Groups A, C, and D.

We identified several candidate genes associated with *P. salmonis* resistance (n=120), which were present in at least one of the groups described previously. These genes are associated with the following biological processes: dependence on kinase activity, GTP hydrolysis, helicase activity, lipid metabolism, cytoskeletal dynamics, and inflammation. To rank the genes, we scored them based on the counting of each of them across following categories: i) species (CS, RT, and AS), ii) trait definitions (DD and BS), and iii) groups (A–D); thus, the maximum score for one particular gene was equal to 9. The prioritized functional candidate genes based on the score described above are shown in **Table 2** and the complete list of unique candidate genes (n=120) can be found in **Supplementary Table S1. sheet: Candidate genes.**

TABLE 2 | Summary of candidate genes associated with *P. salmonis* resistance for CS, RT, and AS ranked by score, which is simply based on the number of appearance of each gene across the following categories: i) species (CS, RT, and AS), ii) trait definitions (DD and BS), and iii) groups (A–D).

Gene symbol	Protein description	Species	Trait	Group	Score[a]
NRTPK	nr-TPK (cytosolic)	CS, RT, and AS	DD and BS	B–D	8
DDX	ATP-dependent RNA helicase DDX	CS and RT	DD	A–C	6
ARL5B	ADP-ribosylation factor protein 5B	CS	DD and BS	A and D	5
LFABP	Fatty acid-binding protein, liver	CS and RT	BS	B and C	5
GIMAP4	GTPase IMAP family member 4	RT	DD and BS	A and D	5
HS3ST3A1	Heparan sulfate glucosamine 3-O-sulfotransferase 3A1	AS	DD and BS	A and D	5
KIF2C	Kinesin protein KIF2C	RT	DD and BS	A and C	5
MYO3B	Myosin-IIIb	CS	DD and BS	A and C	5
NLRP12	NACHT, LRR, and PYD domains-containing protein 12	AS	DD	A, C, and D	5
RAB	Ras-related protein Rab	CS and RT	DD	B and C	5
CACNB2	Voltage-dependent L-type calcium channel subunit beta-2	CS	DD and BS	A and D	5
TDRD9	ATP-dependent RNA helicase TDRD9	CS	DD	A and C	4
FGD	FYVE, RhoGEF, and PH domain-containing protein	RT and AS	DD	B	4
GIMAP8	GTPase IMAP family member 8	RT	DD	A and D	4
HINT1	Histidine triad nucleotide-binding protein 1	AS	DD	C and D	4
KIF15	Kinesin protein KIF15	CS	DD	A and C	4
NLRC3	NACHT, LRR, and CARD domains-containing protein 3	RT	DD	A and D	4
COL12	Collectin-12	CS	BS	D	3
LYZ2	Lysozyme C II	AS	DD	C	3
MRC1	Macrophage mannose receptor 1	CS	BS	D	3
TAPBPR	Tapasin-related protein	RT	DD	D	3

[a] *The maximum score possible for one particular gene is equal to 9.*

DISCUSSION

The comparative genomic strategy used in this study allowed us to identify groups of homologous superfamilies and orthologous genes common to more than one species of salmonids among genes adjacent to SNPs that explain more than 1% of the genetic variance for *P. salmonis* resistance. To our knowledge, this is the first study that aims at identifying and prioritizing functional candidate genes involved in the differential response against bacterial infection by means of comparing results from GWAS mapping across different phylogenetically related salmonid species.

GENETIC ARCHITECTURE OF RESISTANCE TO *P. SALMONIS*

Heritability estimates are in agreement with previous studies aimed to estimate levels of genetic variation for resistance to bacterial diseases in salmonid species. For instance, Vallejo et al. (2016), Vallejo et al. (2017) presented heritabilities ranging from 0.26 to 0.54 and from 0.31 to 0.48 for resistance to bacterial cold water disease in a farmed RT population. The levels of genetic variation observed in the current study are consistent or somewhat higher than previous estimates of heritabilities for resistance to *P. salmonis* depending on the species and the trait definition. For instance, previous heritability values for *P. salmonis* resistance estimated based on pedigree information reached a maximum of 0.16, 0.44, and 0.41 for CS, RT, and AS, respectively (Yáñez et al., 2013; Yáñez et al., 2014; Yáñez et al., 2016; Bassini et al., 2019). When heritability for *P. salmonis* resistance was estimated based on genomic information, the maximum values reported previously were 0.39 and 0.62 for AS and RT, respectively (Bangera et al., 2017; Yoshida et al., 2018a).

Our results show evidence of alleles of medium to large effect involved in resistance to *P. salmonis* in CS and RT. In contrast, for AS, our results suggest that if alleles of large effect do exist, they are at such low frequency that they individually explain a small proportion of the variance for resistance to *P. salmonis*. The identification of genomic regions harboring associated SNPs was based on GWAS using the Bayes C approach, which is more suitable for oligogenic traits (Habier et al., 2011). In a few cases, the same SNP was significantly associated with both trait definitions (DD and BS). This could be the result of pleiotropy, closely linked genes [local linkage disequilibrium (LD)], or by a strong correlation between both traits. For example, we observed the same SNP associated with DD and BS in CS (58185_41 and 24601_47) and RT (AX-89926208 and AX-89966072) among the top 10 SNPs explaining most of the genetic variance for the trait.

Based on the LD of the AS population (measured as r^2), the number of SNPs used for AS (~43K) should be enough to cover the entire genome (Barría et al., 2018b). There is a lack of studies aimed at evaluating the LD and population structure of the current farmed RT population. Based on results from a different RT farmed population, at least 20K SNPs are necessary to cover the whole genome (Vallejo et al., 2018). If the LD levels of the present RT population are similar to those reported by Vallejo et al. (2018), the 23K SNPs used here will most likely cover the whole genome. However, this is not the case for CS. Using a high-density SNP array, Rondeau et al. (In preparation) and Barría et al. (2019) suggested that at least 74K SNPs are necessary for whole-genome studies of the current CS population. The small number of SNPs assayed in this study for CS (9389) most likely affected the identification of markers with a moderate to high effect on resistance to *P. salmonis* in this species.

Candidate Proteins Associated With the Resistance to *P. salmonis*

Whereas the complete set of proteins predicted from reference genomes of CS, RT, and AS consisted of 57,592, 58,925, and 97,738, respectively, the proteins neighboring SNPs associated with resistance (range of 1 Mb) represent less than 1% of the different proteomes. The characterization of the complete set of proteins among species established that the most prevalent homologous superfamily was the P-loop_NTPase. However, as this superfamily contains proteins with at least 21 functions (Shalaeva et al., 2018), it is possible that the high frequency of proteins identified from this group was due to the overall high representation in salmonid genomes. For this reason, we retrieved the sequences of 100 randomly selected proteins from the genomes of CS, RT, and AS and classified them into subfamilies (**Supplementary Figure S3**). The results indicate that P-loop_NTPase is not the most prevalent in any of the salmonid species, which suggests that this homologous superfamily is actually enriched in the regions analyzed and is not a consequence of their high representation in CS, RT, and AS genomes.

When traits are polygenic in nature, the identification of genes underlying them is a challenging task and often depends on previous knowledge of the function of genes adjacent to the associated SNPs (Jiang et al., 2014; Bouwman et al., 2018; Robledo et al., 2019). Our strategy was based on identifying orthologous proteins between the salmonid species and families of homologous proteins in the complete set of proteins adjacent to all the SNPs that explained more than 1% of the genetic variance, without searching for a specific function. The identification of genes directly associated with the innate immune response, after applying all the classification criteria, such as LYZ2, MRC1, COL12, and TAPBPR, suggests that our strategy was successful in finding strong functional candidate genes involved in resistance to *P. salmonis*. Interestingly, about 100 genes not classically associated with the immune system were also identified; among which, 17 were part of at least two of the groups described previously and hence are considered strong candidates for being responsible on trait variation (**Table 2**).

Previously, lysozymes have primarily been described as having a bacteriolytic activity against Gram-positive bacteria; however, the expression of LYZ2 has been shown to be induced in a resistant RT line in response to *Flavobacterium psychrophilum* infection (Langevin et al., 2012) and in AS families in response to *P. salmonis* infection (Pulgar et al., 2015), indicating that the transcriptional regulation of this enzyme in salmonids responds to Gram-negative bacterial infection. MRC1 and COL12 are membrane receptors that display several functions associated with innate immunologic defense, particularly in the recognition of carbohydrate structures of pathogens and as phagocytic receptors of bacteria, yeasts, and

other pathogenic microorganisms (Harris et al., 1992; Ma et al., 2015). It has been reported that enhanced infection in human phagocytes with *Francisella tularensis*, a bacterium phylogenetically related to *P. salmonis*, is mediated by MRC1 (Schulert and Allen, 2006), whereas COL12 led to the activation of the alternative pathway of complement *via* association with properdin, a key positive regulator of the pathway by increment of the half-life of the C3 and C5 convertases (Ma et al., 2015). TAPBPR has been described as a second major histocompatibility complex class I-dedicated chaperone essential to providing specificity for T-cell responses against viruses and bacteria (Hermann et al., 2015) and the related protein tapasin has been shown to be induced in monocyte/macrophage in RT by chum salmon reovirus infection (Sever et al., 2014).

Another set of candidate proteins for SRS resistance in the three salmonid species studied are a cluster of cytosolic nr-TPKs. These proteins are a subgroup of the tyrosine kinase family, enzymes that phosphorylate tyrosine residues, and regulate many cellular functions, such as cell growth and survival, apoptosis, cell adhesion, cytoskeleton remodeling, and differentiation (Neet and Hunter, 1996). Although these proteins are not classically related to the response to pathogens, it has been described that the interaction of T- and B-cell antigen receptors with some nr-TPKs is critical to the activation of lymphocytes by an antigen (Sefton and Taddie, 1994). Moreover, some cellular signaling pathways are hijacked by intracellular pathogens, which can subvert protein phosphorylation to control host immune responses and facilitate invasion and dissemination (Haenssler and Isberg, 2011). It has been described that some bacterial effectors are injected into host cells through their secretion systems where they inhibit the Src kinase. In particular, the effector EspJ, an ADP-ribosyltransferase of the bacteria *Escherichia coli* and *Citrobacter rodentium*, regulates multiple host nr-TPKs *in vivo* by ADP-ribosylation, demonstrating that part of its target protein repertoire involves Src kinases such as YES1 and LYN as well as the adapter SYK (Young et al., 2014; Pollard et al., 2018), all of which were identified in this study in CS, RT, and AS. Remarkably, among the candidate genes, we also identified the small ARL5B, suggesting that an adequate regulation of the activity of nr-RTKs by ADP-ribosylation could be critical to combat *P. salmonis* infection.

Other orthologous candidate genes identified in this study encode for proteins RAB1 and RAB18, both members of the GTPase superfamily. GTPases are a large family of hydrolase enzymes that bind and hydrolyze GTP and play an important role in signal transduction, protein translation, control and cellular differentiation, intracellular transport of vesicles, and cytoskeletal reorganization, among other cellular processes (Bourne et al., 1991). Specifically, RAB GTPases constitute a subfamily of small GTPases known as master regulators of intracellular membrane traffic (Stenmark, 2009). As *P. salmonis* drives the formation of host membrane-derived organelles, the development of these *P. salmonis*-containing vacuoles is dependent on the bacterium's ability to usurp the intracellular membrane system of the fish. Furthermore, two orthologous of FGD were identified in RT and AS. These proteins activate CDC42, a GTPase involved in the organization of the actin cytoskeleton and with a role in early contractile events in phagocytes (Ching et al., 2007). As it has been described that the infective process of *P. salmonis* depends on the exploitation of the actin monomers (Ramírez et al., 2015), the identification in this study of candidate genes that encode for cytoskeletal motor proteins (two kinesins and a myosin) highlights their relevance not only for the reorganization of the cytoskeleton but also for its motility and involvement in the development of the infection (Hoyt et al., 1997). Remarkably, two other candidate proteins associated with SRS resistance are also members of the GTPase superfamily, GIMAP4 and GIMAP8. This is a family of proteins abundantly expressed in lymphocytes and whose function is to contribute in the regulation of apoptosis and the maintenance of T-cell numbers in the organism (Yano et al., 2014).

Another group of orthologous genes code for ATP-dependent RNA helicases DDX24 in CS and DDX47 in RT for DD. The ATP-dependent RNA helicase DDX family, also known as DEAD-box helicases, is required for different cellular processes such as transcription, pre-mRNA processing, ribosome biogenesis, nuclear mRNA export, translation initiation, RNA turnover, and organelle function. The protein structure is very similar to viral RNA helicases and to DNA helicases, which suggests that the fundamental activities of these enzymes are similar (Rocak and Linder, 2004). Viruses also use RNA helicases at various stages of their life cycle. Many viruses carry their own helicases to assist with the synthesis of their genome, but others synthesize their genome within the cell nucleus, which tends to exploit cellular helicases and thus do not encode their own. We also identified the ATP-dependent RNA helicase TDRD9, which has not been directly implicated in infection but was differentially expressed in channel catfish in response to *Aeromonas hydrophila* infection (Li et al., 2013). Mechanistic studies of RNA helicases will allow the determination of the precise role of these helicases in the host-pathogen interaction.

The last group of orthologous genes identified code for two L-FABPs in CS and RT for BS. L-FABPs are abundant in hepatocytes and are known to be associated with lipid metabolism. In addition, these proteins are up-regulated in several types of cancer, but their role in infection remains unclear (Ku et al., 2016). Nevertheless, it has been recently reported that serum and urine L-FABP may be a new diagnostic marker for liver damage in patients with both acute and chronic hepatitis C infection (Cakir et al., 2017). Interestingly, in AS challenged with *P. salmonis*, L-FABP was up-regulated in resistant families and simultaneously down-regulated in susceptible families (Pulgar et al., 2015), suggesting a transcriptional regulation in response to *P. salmonis* infection and a putative expression marker of resistance to SRS.

Genes coding NLRP12, CACNB2, HS3ST3A1, and HINT1 were also selected as candidate genes for SRS resistance. NLRP12 and NLRC3 are two cytosolic proteins that share two functional domains (NACHT and LRR). NLRP12 was one of the best ranked genes, adjacent to the leader SNP and adjacent to more than one SNP simultaneously for DD in AS. This protein functions as an attenuating factor of inflammation in monocytes by negative regulation of the nuclear factor-κB (NF-κB) activation (Fata et al., 2013). In murine macrophages, a significant expression increase has been shown in cells infected with the intracellular parasite *Leishmania major* compared to non-infected macrophages (Fata et al., 2013). NLRC3 is also a negative regulator of the innate immune response mediated by the inhibition of Toll-like

receptor-dependent activation of the transcription factor NF-κB (Schneider et al., 2012). The presence of these genes suggests that the control of the inflammatory reaction in response to *P. salmonis* infection could be essential to combat SRS.

To the best of our knowledge, this is the first time that functional candidate genes underpinning resistance to *P. salmonis* are proposed based on a comparative genomics approach comparing GWAS results for the same trait in different fish genus/species. We hypothesize that variations in the sequences of these genes could play important roles in the host response to *P. salmonis* infection, which could be tested through new genetic approaches such as gene editing using CRISPR-Cas9 and used through GS or more traditional selection practices. All this information together can be used to generate better control and treatment measures for one of the most important bacterial diseases affecting salmon aquaculture.

CONCLUSIONS

Although *P. salmonis* resistance has previously been described as a polygenic trait, our comparative genomics approach based on GWAS results for the same trait in different salmonid species allowed us to identify about 100 candidate genes that may explain resistance to *P. salmonis*. Of these, 21 are suggested to be strong functional candidates influencing the trait. These genes are associated with multiple biological processes, including dependence on kinase activity, GTP hydrolysis, helicase activity, lipid metabolism, cytoskeletal dynamics, inflammation, and the innate immune response. We hypothesize that variations in the sequences of these genes could play an important role in the expression and/or activity of their encoded proteins and consequently in the resistance to *P. salmonis*. This information could be used to generate better control and treatment measures, based on selective breeding or new drug development, for one of the most important bacterial diseases affecting salmon aquaculture.

DATA AVAILABILITY

Genotype and phenotype data generated for this study are available as supplementary material. CS, RT, and AS data are found on Supplementary files 2 to 4, respectively.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All experimental challenges and sampling procedures were approved by the Comité de Bioética Animal from the Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile (Certificate N08-2015).

AUTHOR CONTRIBUTIONS

JY conceived of and designed the study and drafted the manuscript. GY assessed the GWAS analyses. AP and RP designed and assessed the comparative genomic analyses and contributed in the first draft of the manuscript and discussion. LB, ML, and KCo contributed in the RT and AS sampling, genotyping, and QC. AB performed DNA extraction from CS samples, contributed in the initial draft of the manuscript, and performed library construction. KCh performed ddRAD library construction and assessed the comparative sequences analyses between species. RC and JL contributed in the study design, analyses, and discussion. All authors have reviewed and approved the manuscript.

FUNDING

This project was funded by the U-Inicia grant, from the Vicerrectoria de Investigación y Desarrollo, Universidad de Chile. This work was conceived of under the framework of the grant FONDEF NEWTON-PICARTE (IT14I10100), funded by CONICYT (Government of Chile). This work has been partially supported by Núcleo Milenio INVASAL from Iniciativa Científica Milenio (Ministerio de Economía, Fomento y Turismo, Gobierno de Chile). This research was carried out in conjunction with EPIC4 (Enhanced Production in Coho: Culture, Community, Catch), a project supported by the government of Canada through Genome Canada, Genome British Columbia, and Genome Quebec.

ACKNOWLEDGMENTS

Aguas Claras, Pesquera Antares, and Salmones Chaicas provided the CS, RT, and AS datasets, respectively. GY and RC acknowledge the FAPESP (2014/20626-4; 2015/25232-7) and CNPq (308636/2014-7) for the financial support. AB and KCo acknowledge the National Commission of Scientific and Technologic Research (CONICYT) for the funding through the national Ph.D. funding program. RP acknowledges the CONICYT for the funding through the Fondecyt program (11161083). AB acknowledges the Government of Canada for the funding through the Canada–Chile Leadership Exchange Scholarship. We acknowledges the Centro de Investigación en Alimentos para el Bienestar en el Ciclo Vital (ABCvital) for funding this publication. We also want to thank the World Congress on Genetics Applied to Livestock Production as this work has been partially presented on this conference (Yoshida et al., 2018b).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00665/full#supplementary-material

FIGURE S1 | Manhattan plots for resistance to *P. salmonis* measured as DD in CS, RT, and AS. Y-axis represents the percentage of the GEV by each marker.

FIGURE S2 | Manhattan plots for resistance to *P. salmonis* measured as BS in CS, RT, and AS. Y-axis represents the percentage of the GEV by each marker.

FIGURE S3 | Homologous superfamilies (InterPro) associated with 100 random selected proteins from CS, RT, and AS genomes.

REFERENCES

Antoine, T. E., Yakoub, A., Maus, E., Shukla, D., and Tiwari, V. (2014). Zebrafish 3-O-sulfotransferase-4 generated heparan sulfate mediates HSV-1 entry and spread. *PLoS One* 9 (2), e87302. doi: 10.1371/journal.pone.0087302

Bangera, R., Correa, K., Lhorente, J. P., Figueroa, R., and Yáñez, J. M. (2017). Genomic predictions can accelerate selection for resistance against *Piscirickettsia salmonis* in Atlantic salmon (*Salmo salar*). *BMC Genom.* 18, 121. doi: 10.1186/s12864-017-3487-y

Bassini, L. N., Lhorente, J. P., Oyarzún, M., Bangera, R., Yáñez, J. M., and Neira, R. (2019). Genetic parameters for *Piscirickettsia salmonis* resistance, sea lice (*Caligus rogercresseyi*) susceptibility and harvest weight in rainbow trout (*Oncorhynchus mykiss*). *Aquaculture.* 510, 276–282. doi: 10.1016/j.aquaculture.2019.05.008

Barría, A., Christensen, K. A., Yoshida, G. M., Correa, K., Jedlicki, A., Lhorente, J. P., et al. (2018a). Genomic predictions and genome-wide association study of resistance against *Piscirickettsia salmonis* in coho salmon (*Oncorhynchus kisutch*) using ddRAD sequencing. *G3-Genes Genomes Genet.* 8 (4), 1183–1194. doi: 10.1101/124099

Barría, A., Lopez, M. E., Yoshida, G., Carvalheiro, R., and Yáñez, J. M. (2018b). Population genomic structure and genome-wide linkage disequilibrium in farmed Atlantic salmon (*Salmo salar L.*) using dense SNP genotypes. *Front. Genet.* 9, 649. doi: 10.3389/fgene.2018.00649

Barría, A., Christensen, K. A., Yoshida, G., Jedlicki, A., Lhorente, J. P., Davidson, W. S., et al. (2019). Whole genome linkage disequilibrium and effective population size in a coho salmon (*Oncorhynchus kisutch*) breeding population using high density SNP array. *Front. Genet.* 10, 498. doi: 10.3389/fgene.2019.00498

Bishop, S. C., and Woolliams, J. A. (2014). Genomics and disease resistance studies in livestock. *Livest. Sci.* 166, 190–198. doi: 10.1016/j.livsci.2014.04.034

Bourne, H., Sanders, D. A., and Frank McCormick, F. (1991). The GTPase superfamily: conserved structure and molecular mechanism. *Nature* 349, 117–127. doi: 10.1038/349117a0

Bouwman, A. C., Daetwyler, H. D., Chamberlain, A. J., Ponce, C. H., Sargolzaei, M., Schenkel, F. S., et al. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* 50, 362–367. doi: 10.1038/s41588-018-0056-5

Cakir, O. O., Toker, A., Ataseven, H., Demir, A., and Polat, H. (2017). The importance of liver-fatty acid binding protein in diagnosis of liver damage in patients with acute hepatitis. *J. Clin. Diagn. Res.* 11 (4), OC17–OC21. doi: 10.7860/JCDR/2017/24958.9621

Camusetti, M. A., Gallardo, A., Aguilar, D., and Larenas, J. (2015). Análisis de los costos por la utilización de quimioterápicos y vacunas en la salmonicultura. *Salmonexpert.*

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J. H., and De Koning, D.-J. (2011). Stacks: building and genotyping loci *de novo* from short-read sequences. *G3-Genes Genomes Genet.* 1, 171–182. doi: 10.1534/g3.111.000240

Ching, K. H., Kisailus, A. E., and Burbelo, P. D. (2007). Biochemical characterization of distinct regions of SPEC molecules and their role in phagocytosis. *Exp. Cell Res.* 313, 10–21. doi: 10.1016/j.yexcr.2006.09.011

Correa, K., Lhorente, J., Lopez, M., Bassini, L., Naswa, S., Deeb, N., et al. (2015). Genome-wide association analysis reveals loci associated with resistance against *Piscirickettsia salmonis* in two Atlantic salmon (*Salmo salar L.)* chromosomes. *BMC Genom.* 16, 854. doi: 10.1186/s12864-015-2038-7

Cvitanich, J., Garate, O., and Smith, C. E. (1991). The isolation of a rickettsia-like organism causing disease and mortality in Chilean salmonids and its confirmation by Koch's postulate. *J. Fish Dis.* 14, 121–146. doi: 10.1111/j.1365-2761.1991.tb00584.x

Fata, A., Mahmoudian, M., Varasteh, A., and Sankian, M. (2013). Monarch-1 activation in murine macrophage cell line (J774 A.1) infected with Iranian strain of *Leishmania major. Iran. J. Parasitol.* 8 (2), 207–211.

Fernando, R. L., and Garrick, D. (2013). Bayesian methods applied to GWAS. *Genome-wide association studies and genomic prediction*, 1019, 237–274. doi: 10.1007/978-1-62703-447-0_10

Fryer, J. L., and Hedrick, R. P. (2003). *Piscirickettsia salmonis*: a Gram-negative intracellular bacterial pathogen of fish. *J. Fish Dis.* 26, 251–262. doi: 10.1046/j.1365-2761.2003.00460.x

Goddard, M., and Hayes, B. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10, 381–391. doi: 10.1038/nrg2575

Goddard, M. E., Wray, N. R., Verbyla, K., and Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. Stat. Sci. 24, 517–529. doi: 10.1214/09-STS306.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 2008 (10), 3420–3435. doi: 10.1093/nar/gkn176

Habier, D., Fernando, R. L., Kizilkaya, K., Garrick, D. J., Meuwissen, T., Hayes, B., et al. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinform.* 12, 186. doi: 10.1186/1471-2105-12-186

Haenssler, E., and Isberg, R. R. (2011). Control of host cell phosphorylation by *Legionella pneumophila. Front. Microbiol.* 2, 64. doi: 10.3389/fmicb.2011.00064

Harris, N., Super, M., Rits, M., Chang, G., and Ezekowitz, R. A. (1992). Characterization of the murine macrophage mannose receptor: demonstration that the downregulation of receptor expression mediated by interferon-gamma occurs at the level of transcription. *Blood.* 80 (9), 2363–2373.

Hermann, C., Trowsdale, J., and Boyle, L. H. (2015). TAPBPR: a new player in the MHC class I presentation pathway. *Tissue Antigens* 85 (3), 155–166. doi: 10.1111/tan.12538

Hjeltnes, B., Bang-Jensen, B., Bornø, G., Haukaas, A., and Walde, C. S. (Ed.) (2018). The Health Situation in Norwegian Aquaculture 2017. Norwegian Veterinary Institute 108 p.

Houston, R. D., Haley, C. S., Hamilton, A., Guy, D. R., Tinch, A. E., Taggart, J. B., et al. (2008). Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*). *Genetics* 178, 1109–1115. doi: 10.1534/genetics.107.082974

Houston, R. D., Taggart, J. B., Cézard, T., Bekaert, M., Lowe, N. R., Downing, A., et al. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). *BMC Genom.* 15, 90. doi: 10.1186/1471-2164-15-90

Hoyt, M., Hyman, A. A., and Bähler, M. (1997). Motor proteins of the eukaryotic cytoskeleton. *PNAS* 94 (24), 12747–12748. doi: 10.1073/pnas.94.24.12747

Jiang, L., Liu, X., Yang, J., Wang, H., Jiang, J., Liu, L., et al. (2014). Targeted resequencing of GWAS loci reveals novel genetic variants for milk production traits. *BMC Genom.* 15 (1), 1105. doi: 10.1186/s12863-014-0125-4

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109

Ku, C. Y., Liu, Y. H., Lin, H. Y., Lu, S. C., and Lin, J. Y. (2016). Liver fatty acid-binding protein (L-FABP) promotes cellular angiogenesis and migration in hepatocellular carcinoma. *Oncotarget* 7 (14), 18229–18246. doi: 10.18632/oncotarget.7571

Langevin, C., Blanco, M., Martin, S. A., Jouneau, L., Bernardet, J. F., Houel, A., et al. (2012). Transcriptional responses of resistant and susceptible fish clones to the bacterial pathogen *Flavobacterium psychrophilum. PLoS One* 7 (6), e39126. doi: 10.1371/journal.pone.0039126

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinform.* 28, 12:124. doi: 10.1186/1471-2105-12-124

Lee, S. H., Choi, B. H., Lim, D., Gondro, C., Cho, Y. M., Dang, C. G., et al. (2013). Genome-wide association study identifies major loci for carcass weight on BTA14 in Hanwoo (Korean cattle). *PLoS One* 8 (10), e74677. doi: 10.1371/journal.pone.0074677

Legarra, A., Ricard, A., and Filangi, O. (2013). GS3 software package and documentation. Accessed Jan. 2, 2017. http://snp.toulouse.inra.fr/~alegarra

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533 (7602), 200. doi: 10.1038/nature17164

Li, C., Wang, R., Su, B., Luo, Y., Terhune, J., Beck, B., et al. (2013). Evasion of mucosal defenses during *Aeromonas hydrophila* infection of channel catfish (*Ictalurus punctatus*) skin. *Dev. Comp. Immunol.*, 39(4), 447–455. doi: 10.1016/j.dci.2012.11.009

Ma, Y. J., Hein, E., Munthe-Fog, L., Skjoedt, M. O., Bayarri-Olmos, R., Romani, L., et al. (2015). Soluble collectin-12 (CL-12) is a pattern recognition molecule initiating complement activation *via* the alternative pathway. *J. Immunol.* 195 (7), 3365–3373. doi: 10.4049/jimmunol.1500493

Macqueen, D. J., Primmer, C. R., Houston, R. D., Nowak, B. F., Bernatchez, L., Bergseth, S., et al. (2017). Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. *BMC Genom.* 18, 1–9. doi: 10.1186/s12864-017-3862-8

Mandakovic, D., Glasner, B., Maldonado, J., Aravena, P., González, M., Cambiazo, V., et al. (2016). Genomic-based restriction enzyme selection for specific detection of *Piscirickettsia salmonis* by 16S rDNA PCR-RFLP. *Front. Microbiol.* 7, 643. doi: 10.3389/fmicb.2016.00643

Mitchell, A., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi: 10.1093/nar/gky1100

Moen, T., Baranski, M., Sonesson, A. K., and Kjøglum, S. (2009). Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*): population-level associations between markers and trait. *BMC Genom.* 10, 368. doi: 10.1186/1471-2164-10-368

Moen, T., Torgersen, J., Santi, N., Davidson, W. S., Baranski, M., Ødegård, J., et al. (2015). Epithelial cadherin determines resistance to infectious pancreatic necrosis virus in Atlantic salmon. *Genetics* 115, 175406. doi: 10.1534/genetics.115.175406

Neet, K., and Hunter, T. (1996). Vertebrate non-receptor protein-tyrosine kinase families. *Genes Cells* 1 (2), 147–169. doi: 10.1046/j.1365-2443.1996.d01-234.x

Palti, Y., Gao, G., Liu, S., Kent, M. P., Lien, S., Miller, M. R., et al. (2015). The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Mol. Ecol. Resour.* 15, 662–672. doi: 10.1111/1755-0998.12337

Pearse, D., Barson, N., Nome, T., Gao, G., Campbell, M., Abadía-Cardoso, A., et al. (2018) Sex-dependent dominance maintains migration supergene in rainbow trout. *bioRxiv* 504621. doi: 10.1101/504621

Peters, S. O., Kizilkaya, K., Garrick, D. J., Fernando, R. L., Reecy, J. M., Weaber, R. L., et al. (2012). Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in *Brangus heifers. J. Anim. Sci.* 90, 3398–3409. doi: 10.2527/jas.2011-4507

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* 7 (5), e37135. doi: 10.1371/journal.pone.0037135

Pollard, D. J., Berger, C. N., So, E. C., Yu, L., Hadavizadeh, K., Jennings, P., et al. (2018). Broad-spectrum regulation of nonreceptor tyrosine kinases by the bacterial ADP-ribosyltransferase EspJ. *mBio* 9, 2) e00170–18. doi: 10.1128/mBio.00170-18

Pulgar, R., Hödar, C., Travisany, D., Zúñiga, A., Domínguez, C., Maass, A., et al. (2015). Transcriptional response of Atlantic salmon families to *Piscirickettsia salmonis* infection highlights the relevance of the iron-deprivation defence system. *BMC Genom.* 16, 495. doi: 10.1186/s12864-015-1716-9

Ramírez, R., Ǵomez, F. A., and Marshall, S. H. (2015). The infection process of *Piscirickettsia salmonis* in fish macrophages is dependent upon interaction with host-cell clathrin and actin. *FEMS Microbiol. Lett.* 362, 1–8. doi: 10.1093/femsle/fnu012

Robledo, D., Gutiérrez, A., Barría, A., Lhorente, J. P., Houston, R., and Yáñez, J. M. (2019). Discovery and functional annotation of quantitative trait loci affecting resistance to sea lice in Atlantic salmon. *Front. Genet.* 10, 56. doi: 10.3389/fgene.2019.00056

Rozas, M., and Enríquez, R. (2014). *Piscirickettsiosis* and *Piscirickettsia salmonis* in fish: a review. *J. Fish Dis.* 37, 163–188. doi: 10.1111/jfd.12211

Rocak, S., and Linder, P. (2004). DEAD-box proteins: the driving forces behind RNA metabolism. *Nat. Rev. Mol. Cell. Biol.* 5, 232–241. doi: 10.1038/nrm1335

Santana, M. H. A., Junior, G. A. O., Cesar, A. S. M., Freua, M. C., Gomes, R. C., Silva, S. L., et al. (2016). Copy number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. *J. Appl. Genet.* 57, 495–504. doi: 10.1007/s13353-016-0344-7

Schneider, M., Zimmermann, A. G., Roberts, R. A., Zhang, L., Swanson, K. V., Wen, H., et al. (2012). The innate immune sensor NLRC3 attenuates Toll-like receptor signaling *via* modification of the signaling adaptor TRAF6 and transcription factor NF-κB. *Nat. Immunol.* 13 (9), 823–831. doi: 10.1038/ni.2378

Schulert, G. S., and Allen, L. A. (2006). Differential infection of mononuclear phagocytes by *Francisella tularensis*: role of the macrophage mannose receptor. *J. Leukoc. Biol.* 80, 563–571. doi: 10.1189/jlb.0306219

Sefton, B. M., and Taddie, J. A. (1994). Role of tyrosine kinases in lymphocyte activation. *Curr. Opin. Immunol.* 6 (3), 372–379. doi: 10.1016/0952-7915(94)90115-5

Sernapesca (2018). Informe sanitario de salmonicultura en centros marinos. *Primer Semestre*, 2018.

Sever, L., Vo, N. T. K., Bols, N. C., and Dixon, B. (2014). Expression of tapasin in rainbow trout tissues and cell lines and up regulation in a monocyte/macrophage cell line (RTS11) by a viral mimic and viral infection. *Dev. Comp. Immunol.* 44, 86–93. doi: 10.1016/j.dci.2013.11.019

Shalaeva, D. N., Cherepanov, D. A., Galperin, M. Y., Golovin, A. V., and Mulkidjanian, A. Y. (2018). Evolution of cation binding in the active sites of P-loop nucleoside triphosphatases in relation to the basic catalytic mechanism. *Elife* 7, e37373. doi: 10.7554/eLife.37373

Soderlund, C., Bomhoff, M., and Nelson, W. (2011). SyMap v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 39 (10), e68. doi: 10.1093/nar/gkr123

Stenmark, H. (2009). Rab GTPases as coordinators of vesicle traffic. *Nat. Rev. Mol. Cell Biol.* 10, 513–525. doi: 10.1038/nrm2728

Vallejo, R. L., Leeds, T. D., Fragomeni, B. O., Gao, G., Hernandez, A. G., Misztal, I., et al. (2016). Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. *Front. Genet.* 7, 1–13. doi: 10.3389/fgene.2016.00096

Vallejo, R., Liu, S., Gao, G., Fragomeni, B. O., Hernandez, A. G., Leeds, T. D., et al. (2017). Similar genetic architecture with shared and unique quantitative trait loci for bacterial cold water disease resistance in two rainbow trout breeding populations. *Front. Genet.* 8, 1–15. doi: 10.3389/fgene.2017.00156

Vallejo, R. L., Silva, R. M. O., Evenhuis, J. P., Gao, G., Sixin, L., Parsons, J. E., et al. (2018). Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: evidence that long-range LD is a major contributing factor. *J. Anim. Breed. Genet.* 135, 263–274. doi: 10.1111/jbg.12335

Varona, L., García-Cortés, L., and Pérez-Enciso, M. (2001). Bayes factors for detection of quantitative trait loci. *Genet. Sel. Evol.* 33, 133. doi: 10.1186/1297-9686-33-2-133

Vidal, O., Noguera, J. L., Amills, M., Varona, L., Gil, M., Jiménez, N., et al. (2005). Identification of carcass and meat quality quantitative trait loci in a Landrace pig population selected for growth and leanness. *J. Anim. Sci.* 83, 293. doi: 10.2527/2005.832293x

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* 33, 79–86. doi: 10.1002/gepi.20359

Wolc, A., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Dekkers, J. C. M., et al. (2016). Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions. *J. Anim. Sci. Biotechnol.* 7, 7. doi: 10.1186/s40104-016-0066-z

Yano, K., Carter, C., Yoshida, N., Abe, T., Yamada, A., Nitta, , et al. (2014). Gimap3 and Gimap5 cooperate to maintain T-cell numbers in the mouse. *Eur. J. Immunol.* 44 (2), 561–572. doi: 10.1002/eji.201343750

Yáñez, J. M., Bangera, R., Lhorente, J. P., Barría, A., Oyarzún, M., Neira, R., et al. (2016). Negative genetic correlation between resistance against *Piscirickettsia salmonis* and harvest weight in coho salmon (*Oncorhynchus kisutch*). *Aquaculture* 459, 8–13. doi: 10.1016/j.aquaculture.2016.03.020

Yáñez, J. M., Bangera, R., Lhorente, J. P., Oyarzún, M., and Neira, R. (2013). Quantitative genetic variation of resistance against *Piscirickettsia salmonis* in Atlantic salmon (*Salmo salar*). *Aquaculture* 414–415, 155–159. doi: 10.1016/j.aquaculture.2013.08.009

Yáñez, J. M., Lhorente, J. P., Bassini, L. N., Oyarzún, M., Neira, R., and Newman, S. (2014). Genetic co-variation between resistance against both *Caligus rogercresseyi*

and *Piscirickettsia salmonis*, and body weight in Atlantic salmon (*Salmo salar*). *Aquaculture* 433, 295–298. doi: 10.1016/j.aquaculture.2014.06.026

Yañez, J. M., Naswa, S., Lopez, M. E., Bassini, L., Correa, K., Gilbey, J., et al. (2016). Genomewide single nucleotide polymorphism discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. *Mol. Ecol. Resour.* 16, 1002–1011. doi: 10.1111/1755-0998.12503

Yoshida, G. M., Bangera, R., Carvalheiro, R., Correa, K., Figueroa, R., Lhorente, J. P., et al. (2018a). Genomic prediction accuracy for resistance against *Piscirickettsia salmonis* in farmed rainbow trout. *G3-Genes Genomes Genet.* 8, 719–726. doi: 10.1534/g3.117.300499

Yoshida, G. M., Carvalheiro, R., Lhorente, J. P., Correa, K., Barria, A., Figueroa, R. et al., (2018b). "Bayesian genome-wide association analyses reveal different genetic architecture of *Piscirickettsia salmonis* resistance in three salmonid species," in *Proceedings of the 11th World Congress on Genetics Applied to Livestock Production*, (Auckland).

Yoshida, G. M., Lhorente, J. P., Carvalheiro, R., and Yáñez, J. M. (2017). Bayesian genome-wide association analysis for body weight in farmed Atlantic salmon (*Salmo salar L.*). *Anim. Genet.* 48, 698–703. doi: 10.1111/age.12621

Young, J. C., Clements, A., Lang, A. E., Garnett, J. A., Munera, D., Arbeloa, A., et al. (2014). The *Escherichia coli* effector EspJ blocks Src kinase activity *via* amidation and ADP ribosylation. *Nat. Commun.* 5, 5887. doi: 10.1038/ncomms6887

Conflict of Interest Statement: JL and KCo were employed by Benchmark Genetics Chile during the course of the study. The remaining 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 August 2019
doi: 10.3389/fgene.2019.00699

Discovery of Genomic Characteristics and Selection Signatures in Korean Indigenous Goats Through Comparison of 10 Goat Breeds

Jae-Yoon Kim[1,2†], Seongmun Jeong[1†], Kyoung Hyoun Kim[1,2], Won-Jun Lim[1,2], Ho-Yeon Lee[1,2], and Namshin Kim[1,2]*

[1] Genome Editing Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, South Korea, [2] Department of Bioinformatics, KRIBB School of Bioscience, University of Science and Technology (UST), Daejeon, South Korea

OPEN ACCESS

Edited by:
Maria Saura, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Spain

Reviewed by:
Eugenio López-Cortegano, University of Vigo, Spain
Maja Ferenčaković, University of Zagreb, Croatia

***Correspondence:**
Namshin Kim
deepreds@kribb.re.kr

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *23 November 2018*
Accepted: *03 July 2019*
Published: *08 August 2019*

Citation:
Kim J-Y, Jeong S, Kim KH, Lim W-J, Lee H-Y and Kim N (2019) Discovery of Genomic Characteristics and Selection Signatures in Korean Indigenous Goats Through Comparison of 10 Goat Breeds. Front. Genet. 10:699. doi: 10.3389/fgene.2019.00699

Indigenous breeds develop their own genomic characteristics by adapting to local environments or cultures over long periods of time. Most of them are not particularly productive in commercial terms, but they have abilities to survive in harsh environments or tolerate to specific diseases. Their adaptive characteristics play an important role as genetic materials for improving commercial breeds. As a step toward this goal, we analyzed the genome of Korean indigenous goats within 10 goat breeds. We collected 136 goat individuals by sequencing 46 new goats and employing 90 publicly available goats. Our whole-genome data was comprised of three indigenous breeds (Korean indigenous goat, Iranian indigenous goat, and Moroccan indigenous goat; n = 29, 18, 20), six commercial breeds (Saanen, Boer, Anglo-Nubian, British Alpine, Alpine, and Korean crossbred; n = 16, 11, 5, 5, 2, 13), and their ancestral species (*Capra aegagrus*; n = 17). We identified that the Iranian indigenous goat and the Moroccan indigenous goat have relatively similar genomic characteristics within a large category of genomic diversity but found that the Korean indigenous goat has unique genomic characteristics distinguished from the other nine breeds. Through population analysis, we confirmed that these characteristics have resulted from a near-isolated environment with strong genetic drift. The Korean indigenous goat experienced a severe genetic bottleneck upon entering the Korean Peninsula about 2,000 years ago, and has subsequently rarely experienced genetic interactions with other goat breeds. From selection analysis and gene-set enrichment analysis, we revealed selection signals for *Salmonella* infection and cardiomyopathy in the genome of the Korean indigenous goat. These adaptive characteristics were further identified with genomic-based evidence. We uncovered genomic regions of selective sweeps in the LBP and BPI genes (*Salmonella* infection) and the TTN and ITGB6 genes (cardiomyopathy), among several candidate genes. Our research presents unique genomic characteristics and distinctive selection signals of the Korean indigenous goat based on the extensive comparison. Although the adaptive traits require further validation through biological

experiments, our findings are expected to provide a direction for future biodiversity conservation strategies and to contribute another option to genomic-based breeding programmes for improving the viability of *Capra hircus*.

Keywords: Korean indigenous goats, selection signature, genomic characteristics, population genetics, *Capra hircus* (goat)

INTRODUCTION

Goats (*Capra hircus*) are one of the oldest domesticated animals, originating from the wild bezoar goat (*Capra aegagrus*) near the Fertile Crescent of western Asia (Iranian region) (Zeder and Hesse, 2000; Zeder, 2005). Their domestication occurred around the Neolithic period, approximately 10,000 years ago, when human lifestyles moved from hunting to farming (Li and Zhang, 2009). At this time, the goats started supplying milk, meat, fur, and hair to humans in a stable manner, and gradually began to establish a close relationship economically, culturally, and religiously with human civilization (Naderi et al., 2008). As their contribution to humanity increased, goats spread rapidly to the rest of the world following human migration and trade routes (Taberlet et al., 2008; Tresset and Vigne, 2011), and they now comprise more than 1,006 million individuals and over 300 breeds, including commercial and indigenous breeds (http://faostat3.fao.org/browse/Q/QA/E).

Indigenous breeds have locality-specific characteristics, with considerable regional diversity. During the geographical expansion, goats have spread to a wide range of environments spanning hot to cold climates, humid to dry climates, and tropical rainforests to hypoxic high-altitude regions. They have successfully adapted to these diverse environments (Nomura et al., 2013), and have developed distinctive characteristics in their local environments. For instance, in desert areas, one of the Moroccan indigenous goat breeds (the Draa population) has been reported to have acquired the characteristics of frequently gasping to regulate body temperature (Benjelloun et al., 2015). In the highlands, Tibetan indigenous goats have been reported to have developed an oxygen-sensing ability for adapting to hypoxia in high altitudes (Song et al., 2016; Wang et al., 2016). Additionally, Ugandan indigenous goats have been reported to have enhanced their immune competence in order to resist infection by parasites in Africa's tropical environment (Onzima et al., 2018). As useful information, these adaptive characteristics have provided an important base to various breeding programs aimed at improving goat breeds (Giovambattista et al., 2001; Babayan, 2016). For example, Chinese indigenous goats of the Shandong Province, with adaptive characteristics to the humid climate, were used to develop Laoshan dairy goats through selective crossbreeding with Saanen dairy goats (Porter et al., 2016). Due to this breeding effort, the Laoshan dairy goats have acquired both humid climate adaptability (Chinese indigenous goats) and high dairy productivity (Saanen goats) (Li et al., 2008). Also, Indonesian indigenous goats (Katjang goats), which are adapted to the equatorial climate, were utilized to develop Peranakan Etawah goats through crossbreeding with Indian indigenous goats (Jamunapari goats) (Porter et al., 2016). The Peranakan Etawah goats, thus, have shown both equatorial climate adaptability (Indonesian indigenous goats) and high dairy and meat productivity (Indian indigenous goats) (Sodiq, 2004).

Korean indigenous goats (KNG) are the only indigenous goat breed inhabiting the Korean Peninsula. The KNG is characterized by black fur (**Figure 1A**) and is registered with the Food and Agriculture Organization of the United Nations as a single breed (Kim et al., 2011). The origin of the KNG is unclear, but according to previous reports and historical documents, it is estimated that they moved into the Korean Peninsula at least 2,000 years ago after passing through the Northern Mongolia or the Southern coast of China (Tavakolian, 2000; Zeder and Hesse, 2000). Since the influx, the KNG has developed its own unique characteristics while adapting to the peninsula environment for a long time (Rischkowsky and Pilling, 2007). Some of their unique characteristics have been reported through several previous studies. In terms of genetic diversity, Odahara et al. reported that the KNG has not undergone genetic interactions with imported breeds (Odahara et al., 2006). With respect to disease resistance, Jang revealed that *Salmonella* species was not isolated from the feces of either 49 KNG with symptoms of diarrhea or 620 healthy KNG (Jang, 1995). Kang, and Lee et al. also revealed that the KNG lacks *Salmonella* infection due to their excellent antibody production and innate resistance factors (Kang and Tak, 1996; Lee et al., 2000). In addition, Lee et al. reported that the KNG has an adaptive characteristic associated with lumbar paralysis resistance when compared with their crossbreed, Korean crossbred goats (KCB) (Lee et al., 2016). Although the KNG has not been investigated in as much depth as other breeds, these studies have suggested that KNG possesses unique and useful characteristics as an indigenous breed, and also have raised the need for additional research to further reveal their characteristics.

In recent times, the KNG is gradually losing its unique characteristics. After an agreement with the World Trade

Abbreviations: AB, Australian Boer; AN, Anglo-Nubian; ARVC, arrhythmogenic right ventricular cardiomyopathy; AS, Australian Saanen; BA, British Alpine; CA, *Capra aegagrus*; DCM, dilated cardiomyopathy; F, inbreeding coefficient; FA, French Alpine; FS, French Saanen; Fst, fixation index value; GATK, Genome Analysis Toolkit; GSEA, gene-set enrichment analysis; HCM, hypertrophic cardiomyopathy; He, expected heterozygous genotype frequency; Ho, observed heterozygous genotype frequency; ING, Iranian native goat; KB, Korean Boer; KCB, Korean crossbred; KNG, Korean native goat; KS, Korean Saanen; LD, linkage disequilibrium; MNG, Moroccan native goat; N_e, effective population size; PCA, principal component analysis; π, nucleotide diversity; SB, Swiss Boer; SS, Swiss Saanen; XP-CLR, cross-population composite likelihood ratio; XP-EHH, cross-population extended haplotype homozygosity.

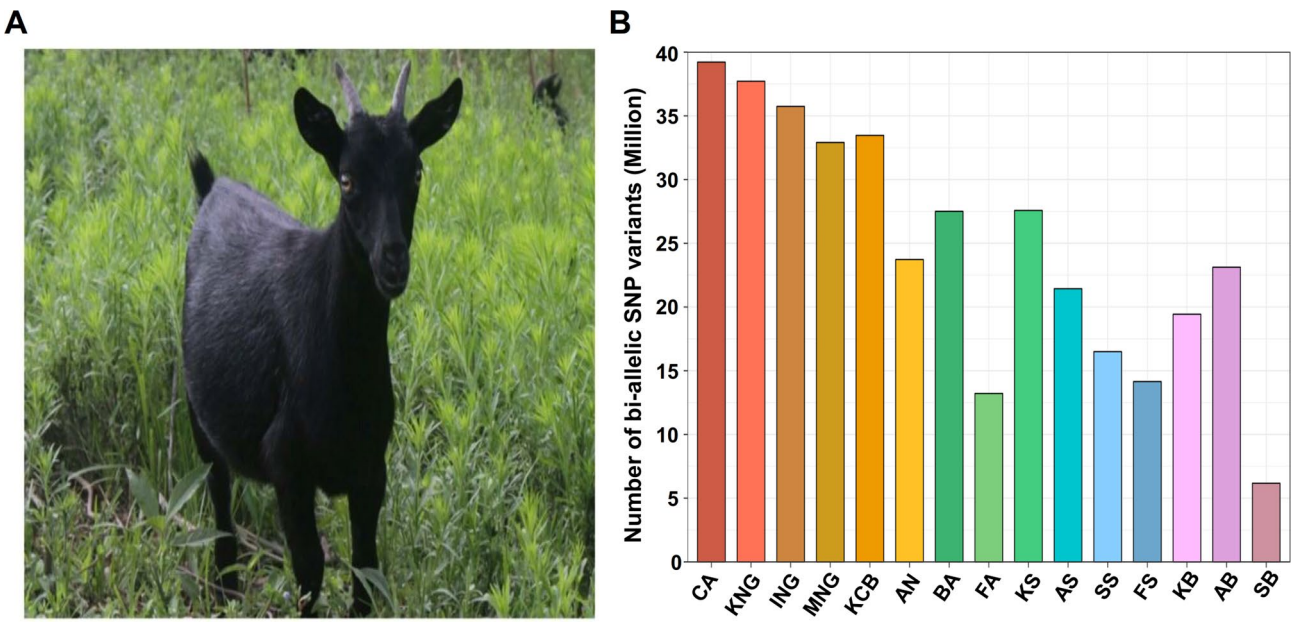


FIGURE 1 | Appearance of Korean indigenous goats (KNG) and distribution of bi-allelic single nucleotide polymorphism (SNP) variants. **(A)** Appearance of KNG with black coat color. **(B)** Distribution of bi-allelic SNPs for 15 goat populations. The *y-axis* represents the number of bi-allelic SNPs detected in each goat population with respect to the reference genome CHIR 2.0. The abbreviation CA means *C. aegagrus* (See **Data sheet 1: Table S5** for a detailed summary of each breed).

Organization in the 1990s, various commercial breeds have been introduced into Korea in earnest (Son, 1999). Since then, the KNG with relatively low commercial productivity has been extensively crossed with imported breeds such as Boer or Saanen, and more recently the KCB, a crossbred of KNG, has even been developed by these hybridizations. However, the genomic characteristics and biodiversity of the KNG have not been extensively investigated. Only one study has reported on the KNG's characteristics at the whole-genome level (Lee et al., 2016). This study compared KNG with only their hybrid KCB, and thus had limitations in identifying the various genomic and adaptive characteristics of KNG. Also, other studies on the KNG's *Salmonella* infection and isolated environmental characteristics require further research based on the genome (Jang, 1995; Kang and Tak, 1996; Lee et al., 2000; Odahara et al., 2006).

In this paper, we conducted a comparative genomic study to reveal the genomic and adaptive characteristics of KNG. For extensive comparison, we analysed whole-genome variations of 10 goat breeds comprising three indigenous breeds (KNG, Iranian indigenous goats, and Moroccan indigenous goats), six commercial breeds (KCB, Saanen, Alpine, British-Alpine, Boer, and Anglo-Nubian), and an ancestral species (*C. aegagrus*). We not only identified the characteristics of KNG, but also established for the first time the genetic relationships between the 10 goat breeds, with the criteria of the ancestral species and Iranian indigenous goats. The aims of our study were: to unravel the genomic characteristics of KNG in the 10 goat breeds; to present genomic evidence that KNG has rarely experienced interactions with other breeds; and to elucidate selection signals that KNG has adapted to their environment.

MATERIAL AND METHODS

Sample Preparation and Re-Sequencing

Blood samples from 46 goats were obtained from the Animal Genetic Resources Station, National Institute of Animal Science, Rural Development Administration in Korea. The blood samples comprised 14 Korean indigenous goats, 10 Korean Saanen, and 4 Korean Boer, which live in Korea; and 5 Anglo-Nubian, 5 British Alpine, 6 Australian Boer, and 2 Australian Saanen, which live in Australia. DNA was isolated according to the manufacturer's protocol using the G-DEXTMIIb Genome DNA Extraction Kit (iNtRoN Biotechnology, Korea), and 3 μg of this genomic DNA was randomly sheared to have an insert size of 300bp using the Covaris System. The fragments of sheared DNA were amplified with the TruSeq DNA Sample Prep Kit (Illumina, USA) and were then sequenced as paired-end reads with approximately 10-fold coverage using the Illumina HiSeq 2000 platform with the TruSeq SBS Kit v3-HS (Illumina). These 46 goat sequences with paired-end reads were deposited in the European Nucleotide Archive under the accession number PRJEB25062. Additionally, we used 90 publicly-available goat genomes comprising 15 Korean indigenous goats, 13 Korean crossbred, 20 Moroccan indigenous goats, 18 Iranian indigenous goats, 17 C. *aegagrus*, two French Saanen, two French Alpine, two Swiss Saanen, and one Swiss Boer. As for the Australia Saanen, the French Saanen, the Swiss Saanen, and the Swiss Boer, we mentioned only their overall trends, because they could have a sampling bias due to a small number of samples. We mainly used the integrated Saanen population (n = 16) and Boer population (n = 11). Additional information of these breeds about sample sizes and bio-project

IDs are summarized in **Data sheet 1: Table S1**, and brief sampling information is provided in **Data sheet 1: Note S1.1**.

Data Processing and Variant Calling

We conducted a per-base sequence quality check for the 136 goat samples using FastQC (Andrews, 2010) and controlled sequences with low quality using NGSQCToolkit (Patel and Jain, 2012). The paired-end sequence reads of each of the 136 samples were then mapped against the reference goat genome, the genome of China's Yunnan black goat 2.0 version (CHIR v2.0), through BWA (Li and Durbin, 2010). The mapped BAM files were sorted into the genomic coordinates of their reference genome using the Picard software's "AddOrReplaceReadGroup" (http://broadinstitute.github.io/picard), and potential PCR duplicates were removed using the "MarkDuplicates" option of the software (**Data sheet 1: Tables S2** and **S3**). Then, the "RealignerTargetCreator" and "IndelRealigner" of the Genome Analysis Toolkit v3.7 (GATK) (Van der Auwera et al., 2013) were used to correct misalignments resulting from INDELs that may exist in the mapped reads. Following this preparation, we generated gVCF files for each of the 136 samples, which were called to all base sites of the reference genome using the GATK's "HaplotypeCaller," combined these gVCF files as one gVCF file through the GATK's "CombineGVCFs," and converted the file into a VCF file using the GATK's "GenotypeVCFs." To exclude as many false positively called variants as possible, the arguments "Variant Filtration" and "Select Variants" of the GATK were adopted with the following options: 1) Phred-scaled quality score (QUAL) < 35.0; 2) Quality score by depth (QD) < 5.0; 3) Genotype quality score (GQ) < 15.0; 4) Mapping quality score (MQ) < 30.0; 5) Phred-scaled *P*-value score of Fisher's exact test for identifying strand bias (FS) > 30.0; 6) Depth of coverage across all samples (DP) < 7; 7) Rank sum test for mapping quality of reference and alternative reads (MQRankSum) < −2.0; and 8) Ranks sum test on the bias of the relative positions of the reference alleles and the alternative alleles in the read (ReadPosRankSum) < −2.0. We additionally filtered variants with genotype missing rates of >50% in order to use relatively common variants. Single nucleotide polymorphism (SNP) and INDEL variants were then separated from the VCF, and bi-allele-type SNPs were extracted (**Figure 1B** and **Data sheet 1: Table S4**). For loci with three or more alleles, we maintained only the allele with the highest allele frequency as the only alternative allele representing the corresponding locus. Lastly, haplotype phasing and imputation were conducted using BEAGLE v4.18 (Browning and Browning, 2007). This variant calling process was also performed for each breed to obtain breed-specific SNPs (**Data sheet 1: Table S5**). The functional effects of these SNPs on the genomic and protein regions were annotated by SnpEff (Cingolani et al., 2012) (**Data sheet 1: Table S6**). Since the gene set of the reference genome CHIR v2.0 has not been fully developed, we used a gene set that mapped the CHIR v1.0 gene set to the CHIR v2.0 reference genome using GMAP (Wu and Watanabe, 2005).

General Genomic Characteristics

Nucleotide diversity (π) was calculated by sliding 50 Kb with a window size of 100 Kb using VCFtools v4.1 (Danecek et al., 2011). Inbreeding coefficient (*F*) was calculated using the same software. The individual's *F* value was obtained by averaging the deviations of observed heterozygous genotype frequency (Ho) from expected heterozygous genotype frequency under random mating (He) ($F = 1 - Ho/He$) for all loci, and the breed's *F* value was derived by averaging these *F* values of all individuals belonging to each breed. Linkage disequilibrium (LD) was measured as r^2 statistic suggested by Hill and Robertson (Hill and Robertson, 1968), and computed using all bi-allelic SNPs through PopLDdecay v3.2 (https://github.com/BGI-shenzhen/PopLDdecay). Then, the averages of pairwise LDs for all SNPs within 30 Kb, 50 Kb, 100 Kb, and 500 Kb regions were calculated. A summary of these three measurements, π, *F*, and LD, is provided in **Data sheet 1: Table S7**, and the average degree of collapse of the LD up to 500 Kb is displayed in **Data sheet 1: Figures S1A–D**.

Population Differentiation and Genetic Structure

Fixation index value (Fst) (Weir and Cockerham, 1984) was calculated for 15 goat populations by sliding 50 Kb with a window size of 100 Kb using the VCFtools (**Data sheet 1: Table S8**). A phylogenetic tree was computed based on the identity-by-state matrix (**Data sheet 1: Figure S2**) which was calculated from all 136 goat samples using Plink v1.90b (Purcell et al., 2007) and reconstructed using the BIO-neighbor-joining algorithm (Gascuel, 1997) which is an improved version of the neighbor-joining algorithm. Then, the tree was visualized using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) (**Figure 2A**). A structure analysis was performed using FAST-STRUCTURE v1.0 (Raj et al., 2014), which is based on a variational Bayesian framework (**Figure 2B** and **Data sheet 1: Figure S3**). The number of genetic clusters (K) was estimated from 2 to 10, and each genetic cluster was calculated *via* cross-validation 10 times with the 1e-7 convergence criterion using the simple prior model. In our case, with the high population structure, the simple prior model was appropriate. A principal component analysis (PCA) was performed by the singular value decomposition of the relationship matrix derived from the Kimura two-parameter model (Kimura, 1980). The PCA plots were displayed using principal components 1, 2, and 3, and the scree plots were presented with their eigenvectors and explanatory powers (**Figures 2C**, **D** and **Data sheet 1: Figures S4A–D**).

Inference of Gene Flow and Demographic History

A maximum likelihood tree indicative of the genetic relationships among populations with directions of genetic drift and gene flow was reconstructed using TreeMix v1.13 (Pickrell and Pritchard, 2012) (**Figure 3A** and **Data sheet 1: Figure S5**). *C. aegagrus* was used as the root, and the block size for estimating the covariance matrix was chosen as 200 Kb, in consideration of the LD. The number of migration events was calculated as six, considering the complexity of our goat populations. The scale bar in the upper left corner represents the standard error of the tree, which represents the variation width of the tree estimated from the 10-time calculations. The reliability of this maximum likelihood

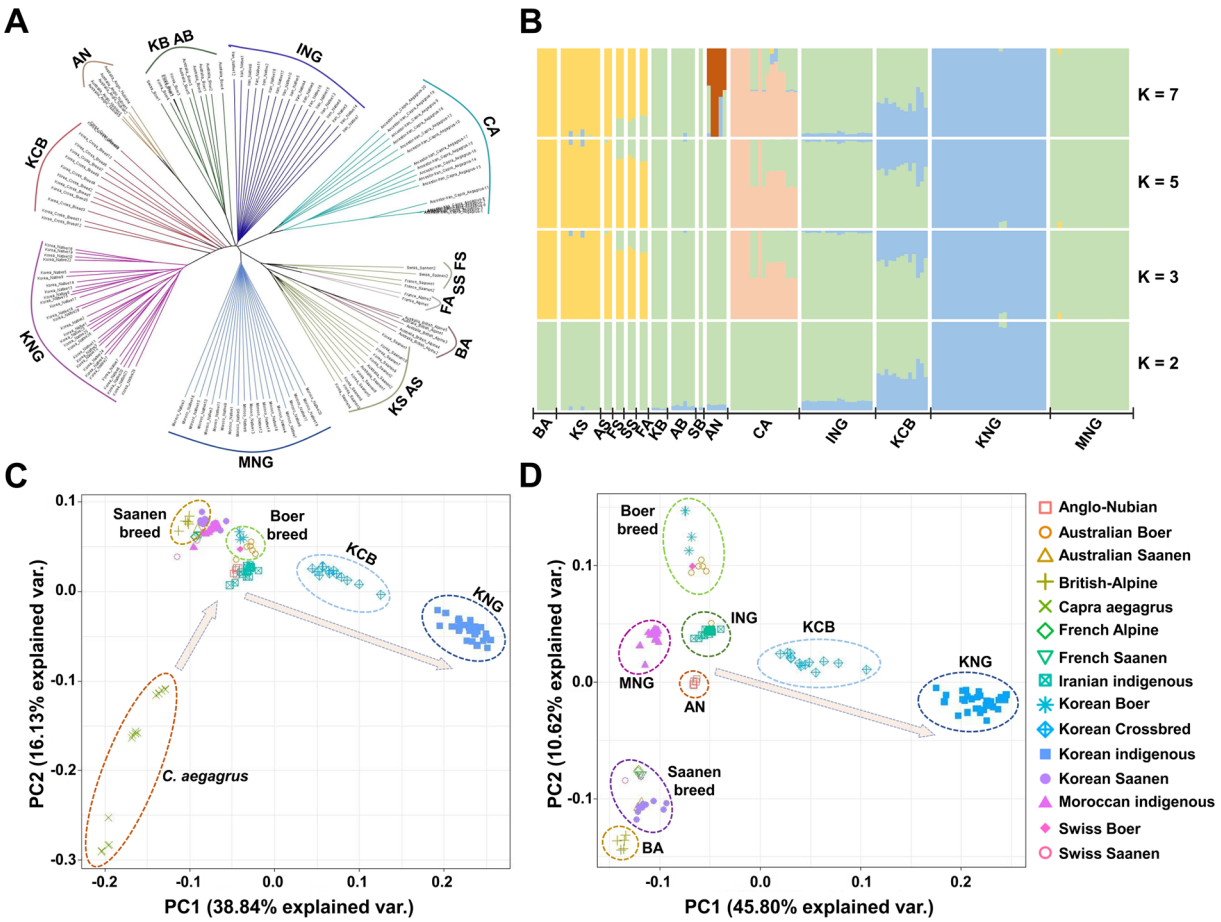


FIGURE 2 | Genomic structure and relationship of 15 goat populations. **(A)** Bio-neighbor-joining tree for the relationship computed using the identical-by-state (IBS). **(B)** Genomic structure computed based on the maximum likelihood-based clustering algorithm. **(C), (D)** Principal component analysis for the genomic differentiation including or excluding *C. aegagrus*, respectively.

tree was confirmed *via* multiple repeats (**Data sheet 1: Figures S6A–B**). To further validate the migration edges identified in the maximum likelihood tree, we conducted the Patterson's D-statistic test (Durand et al., 2011) and the 3-population test (Reich et al., 2009) (**Data sheet 2**). The demographic history of each population was estimated using PopSizeABC (Boitard et al., 2016) (**Figure 3B** and **Data sheet 1: Figure S7**). The mutation rate of a base per generation was calculated as 1e-8, and the lower and upper bounds of the recombination rate were calculated as 1e-9 and 1e-8, respectively. The criterion of minor allele frequency was less than 0.2, and the segment size was 2,000,000. These demographic estimates were obtained through 100,000 iterations (**Data sheet 1: Figure S8** and **Table S9**).

Detection of Selection Signals and Gene-Set Enrichment Analysis

Cross-population extended haplotype homozygosity (XP-EHH) and cross-population composite likelihood ratio (XP-CLR) methods were analyzed using Selscan v1.1.0b (Szpiech and Hernandez, 2014) and XP-CLR v1.0 (Chen et al., 2010), respectively. The genetic positions were assumed to be equivalent to the physical positions due to the absence of a genetic map (1Mb = 1cM). The raw scores of the XP-EHH were standardized to the mean and the standard deviation, and $-\log(1 - p\text{-value})$ of the two-tailed test was calculated through the empirical distribution (**Data sheet 1: Figure S9**). Based on this *p*-value, the outlier regions belonging to the top 0.1% were classified into specific candidate regions for further analysis, and the genes closest to these regions were designated as putative selected genes (**Data sheet 3**). The XP-CLR analysis was calculated by sliding 5 Kb with a window size of 10 Kb. A maximum of 2,000 SNPs were considered for each window, and a correlation level of 0.95 was used. The outlier regions with the top 0.1% of the raw scores were regarded as putative selection regions (**Data sheet 1: Figure S10**), and the closest genes to these candidate regions were designated as selected putative genes (**Data sheet 1: Figures S11A–J** and **Data sheet 4**). The number of selected genes detected in 10 goat populations by these two methods is summarized in **Data sheet 1: Table S10**. To identify the patterns of the adaptation process, we pooled the candidate genes detected by the XP-CLR and XP-EHH methods into one gene set for each population and performed gene set enrichment analysis

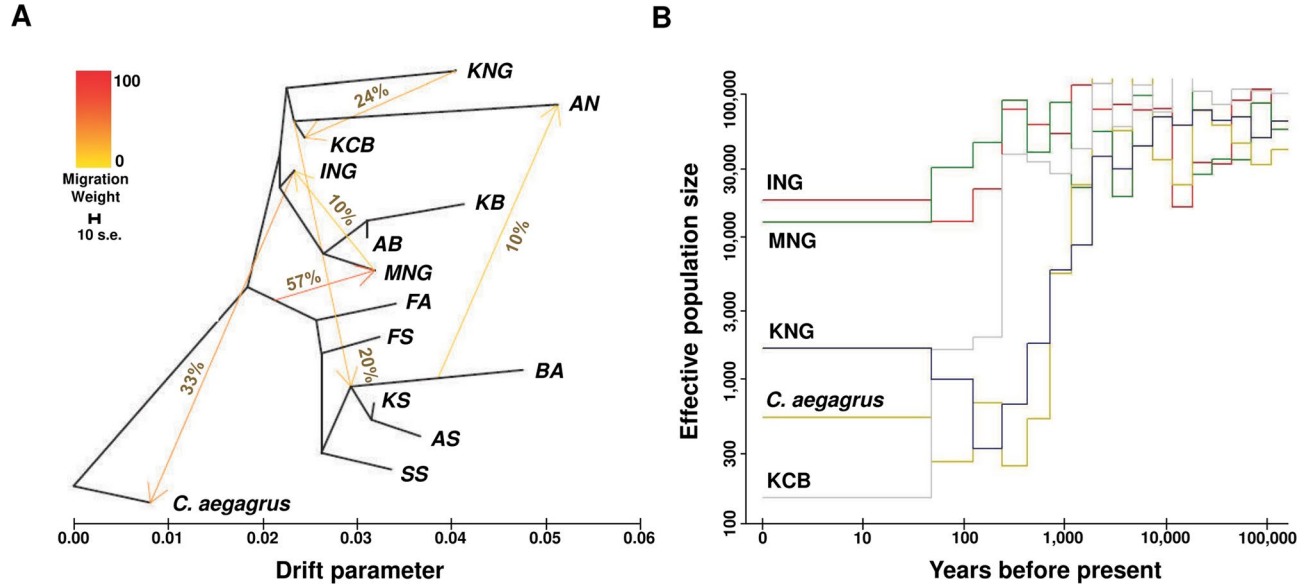


FIGURE 3 | Inferred patterns of gene flow, genetic drift, and effective population size. **(A)** Pattern of population splits and mixture among 14 goat populations. The drift parameter is proportional to effective population size. The migration proportion, above the arrow, indicates the fraction of ancestry derived from the source population. The scale bar represents the standard error calculated 10 times from the sample covariance matrix. **(B)** Inferred demographic history for *C. aegagrus*, KCB, and indigenous breeds (ING, MNG, and KNG) (See **Data sheet 1: Figure S7** and **Table S9**).

(GSEA) for each pooled gene set using GeneTrail2 v1.5 (Stöckel et al., 2016) These candidate genes were grouped into various categories involved in similar functions, pathways, and biological processes through the Kyoto Encyclopedia of Genes and Genomes pathway database and the Gene Ontology database. The statistical significance level for the categories was a *p*-value of about 0.05 adjusted using the Benjamini-Hochberg method (**Table 1** and **Data sheet 5**).

RESULTS

Data Collection, Re-Sequencing, and Identification of SNPs and INDELs

We generated whole-genome data for 46 goats and collected publicly available whole genome data for an additional 90 goats (**Data sheet 1: Table S1**). Our whole-genome data of the 136 individual goats covered 10 goat breeds [*C. aegagrus*, Iranian indigenous goats (ING), Moroccan indigenous goats (MNG), Korean indigenous goats (KNG), Korean crossbred (KCB), Saanen, Boer, British Alpine (BA), French alpine (FA) and Anglo-Nubian (AN)]. The Saanen and the Boer breeds constituted four sub-groups [Swiss Saanen (SS), Australian Saanen (AS), Korean Saanen (KS), and French Saanen (FS)] and three sub-groups [Australian Boer (AB), Korean Boer (KB), and Swiss Boer (SB)]. In total, 50.13 billion reads of 136 goat samples were aligned to the goat reference genome CHIR v2.0 (Dong et al., 2013). The average alignment rate was 99.47%, and it covered 98.61% of the reference genome (**Data sheet 1: Tables S2** and **S3**). The average depths of the reads that removed potential PCR duplicates were 13.07X in the 90 publicly available goats and 12.14X in the 46 newly sequenced goats. To exclude as many false-positive called variants as possible, we strictly performed various filtering processes because the depths were not high (see "Materials and Methods").

TABLE 1 | Significantly enriched terms identified in Korean indigenous goats (KNG) through gene set enrichment analysis (GSEA) (see **Data sheet 5** for summary of all enriched terms).

Selected terms	Number of selected genes[a]	Adjusted *p*-value range[b]	Selected breeds[c]
Salmonella infection	11	(0.0170, 0.0648)	CA, IN, MN, TB, KB, TS, KS, AN, BA
Dilated cardiomyopathy (DCM)	21	(0.0002, 0.0318)	CA, IN, MN, KCB, TB, KB, TS, KS, AN
Hypertrophic cardiomyopathy (HCM)	15	(0.0041, 0.0318)	CA, IN, MN, TB, KB, TS, KS, AN
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	17	(0.0001, 0.0575)	CA, IN, MN, TB, KB, TS, KS, AN, BA

[a]*The total number of selected genes enriched in the selected term.*

[b]*The range of the minimum and maximum values of the adjusted p-values which each selected breed has for the selected term.*

[c]*Abbreviations in the selected breeds column means: IN is Iranian native goat, MN is Moroccan native goat, CA is C. aegagrus, KCB is Korean crossbred, KB is Korean Boer, KS is Korean Saanen, BA is British Alpine, AN is Anglo-Nubian, TB is the entire Boer group, and TS is the entire Saanen group.*

After the variant calling and the filtering processes, a total of 5,629,521 INDEL variants and 39,830,354 bi-allelic SNPs were finally identified. Breed-specific SNPs were then extracted from the bi-allelic SNPs (**Figure 1B** and **Data sheet 1: Tables S4** and **S5**). The numbers of bi-allelic SNPs were markedly different between commercial and indigenous breeds (including *C. aegagrus*). In the commercial breeds, the number of bi-allelic SNPs was at least 20% fewer than those of the indigenous breeds and detection in exon regions was also at least 32% less (**Data sheet 1: Table S6**). These lower tendencies in commercial breeds are considered to be the result of efforts to maintain breed homogeneity through artificial selection. One of the indigenous breeds, KNG, showed the highest number of bi-allelic SNPs (37,715,208) and missense mutations (188,265) except for *C. aegagrus*. Considering that the reference genome, China's Yunnan black goat, has the same black coat color as KNG and the origin of KNG is indirectly related to China, these observations suggest that KNG possesses many SNPs that might have a functional influence on the formation of its unique genomic characteristics. From the following analysis, we used 38,658,962 bi-allelic autosomal SNPs, with an average distance of 64.88 bases between SNPs. This data set covered a significant portion of the reference genome. Additional results and discussions for other breeds are provided in **Data sheet 1: Note S1.2**.

General Genomic Characteristics

To obtain a catalog of general genomic characteristics of the 10 goat breeds comprising the 15 goat populations, we estimated nucleotide diversity (π), inbreeding coefficient (F), and linkage disequilibrium (LD) (**Data sheet 1: Table S7**). The three estimates were quite variable between the populations. The π was the highest in ING and KCB, at 0.001908 and 0.001804, respectively, while the F was the highest in *C. aegagrus* and ING, at 0.0682 and 0.0622, respectively. The average LD patterns showed rapid declines within 50 Kb in all populations and, except for AN, BA and KNG, reached a plateau at around 200 Kb, implicating independent haplotype structures (**Data sheet 1: Figures S1A–D**). The average LD up to 500 Kb was the highest in AN and BA, at 0.3275 and 0.2888, respectively. In this catalog, KNG exhibited the distinctive genomic characteristics close to an isolated population. Among the indigenous breeds, the LD pattern of KNG was the highest at 0.0884, while the π and the F were the lowest, at 0.001472 and 0.01661, respectively. The higher LD indicates that KNG had initiated its breeding history with a limited number of founders in which recombination events occurred infrequently (Chakraborty and Deka, 2005), and has formed a comparatively homogeneous genome until now without few external pressures. The reduced π and F also indicate that KNG is a homogeneous population which has a relatively small number of homozygous genotypes. Along with the detection of the largest number of bi-allelic SNPs (**Figure 1B** and **Data sheet 1: Table S5**), these results suggest that KNG possesses many distinctive SNPs formed by their environmental influence. Moreover, the lower π was consistent with a previous study reporting that KNG has a lower genetic diversity than other Asian goat populations (Odahara et al., 2006). Additional results and discussion on the genomic characteristics for other goat breeds are provided in **Data sheet 1: Note S1.3**.

Population Differentiation and Genic Structure

To obtain a refined picture of the 15 goat populations, we examined the patterns of genetic differentiation and genomic structure using reconstructed tree analysis (Gascuel, 1997), structure analysis (Raj et al., 2014), principal component analysis (PCA), and fixation index value (Fst) (Weir and Cockerham, 1984). These analyses revealed that KNG has genomic characteristics distinct from those of other goat populations (**Figures 2A–D**). The reconstructed tree showed that seven goat breeds, except for the Saanen and Alpine breeds, form their own clade which is genetically distinguished from each other (**Figure 2A**). The Saanen and Alpine breeds, improved similarly for the dairy purpose, formed three sister clades within a common large clade. In the structure analysis calculated ranging from $K = 2$ to $K = 10$ (**Figure 2B** and **Data sheet 1: Figure S3**), we obtained the most reasonable biological interpretation at $K = 7$. At $K = 2$, KNG, ING, and KCB were separated with having a common genomic composition (blue color). With increasing K values, *C. aegagrus*, Saanen and Boer breeds were further separated, and at the $K = 7$, AN was lastly separated with highly mixed genomic compositions observed. We found that KNG has almost a single genomic composition that is not mixed with other goat breeds. This finding indicates that a substantial portion of the KNG's genome is distinct from those of other goat breeds. In addition, the result that the KNG's genomic composition coincided with one of ING's, suggests that KNG originated from the Iranian region where *C. hircus* appeared. The PCA clarified the complex stratifications of 15 goat populations. The first PC in **Figure 2C**, explaining 38.84% of the total genetic variation, separated *C. aegagrus* the farthest to the left and KNG the farthest to the right. The second PC, explaining 16.13% of the total genetic variation, separated Boer and Saanen breeds. **Figure 2D**, which excluded the out-group *C. aegagrus*, distinguished this complex structure in more detail. Centered on ING nearest to the wild-type, Boer breeds and Saanen breeds were separated from each other up and down, and then KNG was separated to the rightmost. The KCB, which was formed by hybridization of the KNG with various commercial breeds, was positioned between ING and KNG (**Figure 2D**). The Fst, calculated in a pair-wise manner, supported these qualitative distinctions (**Data sheet 1: Table S8**). The KNG showed the highest differentiation level between the 14 goat populations. The KNG had the highest differentiation level with BA (0.1908), which was the farthest from KNG, and had the lowest differentiation level with KCB (0.0733), which was the nearest to KNG, as shown in **Figures 2A–D**. Our refined picture indicates that KNG has unique genomic characteristics, and it suggests that the KNG has formed its own genome by accumulating the pressure of their local environment for a long period time, with little interaction with other goat populations. Additional results and discussion on the genomic status of other goat populations are provided in **Data sheet 1: Note S1.4**.

Gene Flow and Demographic History

To visualize the genetic interaction of the 14 goat populations (excluding SB), we constructed a maximum likelihood tree using TreeMix (Pickrell and Pritchard, 2012) (**Figure 3A** and **Data sheet 1: Figure S5**). In this dendrogram, *C. hircus* was differentiated from *C. aegagrus* and then largely divided into the dairy breed and the meat type breed. KNG was directly differentiated from *C. aegagrus* and later ING, and showed an independent long branch indicating a high level of genetic drift. We found evidence that *C. aegagrus* and the indigenous breeds (ING and MNG) have interacted with each other, but no evidence that KNG has interacted with other goat breeds, except for KCB. This evidence was also not detected in additional analyses using the D-statistic (Durand et al., 2011) and 3-population (Reich et al., 2009) tests. These two tests supported the hypothesis that KNG has interacted only with KCB (**Data sheet 2**). Our result provides genomic evidence for existing reports that KNG has not gone through any genetic interchanges with imported breeds since its influx into the Korean Peninsula (Son, 1999; Kim et al., 2011).

We inferred the effective population size (N_e) over the past time, in order to clarify the genetic drift which indigenous breeds and *C. aegagrus* have experienced (**Figure 3A**). The amount of genetic drift depends on the N_e (Ewens, 1990; Ballou et al., 2010; Frankham et al., 2010; Gasca-Pineda et al., 2013) (**Figure 3B** and **Data sheet 1: Figures S7A–D**). The reconstructed N_e patterns showed the domestication event between *C. aegagrus* and *C. hircus*, and a demographic event of KNG. The *C. aegagrus* maintained a high N_e for a long time, despite the appearance of *C. hircus*, which was domesticated about 10,000 years ago. However, the N_e started to decrease sharply about 1,000 years ago and has remained low until now. During the same period, the N_e of ING and MNG (both of which belong to *C. hircus*) increased about 1.5 times, and their genetic diversity has been maintained without loss until now. The crossing pattern of these N_e between *C. aegagrus* and *C. hircus* indicates the increased utilization of *C. hircus* and the decreased utilization of *C. aegagrus,* due to the successful domestication of *C. hircus*. Notably, at the time the N_e of these indigenous breeds began to increase, KNG experienced a serious loss of genetic diversity. This period nearly coincided with the time when KNG was estimated to be introduced into the Korean Peninsula (about 2,000 years ago) (Kang, 1967; Son, 1999). Since that time, the N_e of KNG steadily decreased until 100 years ago. At present, the N_e has increased slightly, but it showed still much lower than those for other indigenous breeds (**Data sheet 1: Table S9**). This N_e pattern represents that KNG had experienced a genetic bottleneck event during its influx into the Korea Peninsula and has relatively well adapted to the Korean environment since then. The 90% credible intervals of the estimated N_e for each population are displayed in **Data sheet 1: Figure S8**, and additional results and discussion for other goat populations are provided in **Data sheet 1: Note S1.5**.

Detection of Selection Signals and Selective Sweep Regions in KNG

Nature selects a genomic region associated with specific traits such as disease or parasite resistance and temperature, or high-altitude adaptation, in order to increase the organisms' chance of survival or reproduction in a particular environment (Futuyma, 2009). In the case of an isolated population, their genome is more susceptible to natural selection due to an environment with low confounding effects (Losos and Ricklefs, 2009; Pergams and Lawler, 2009). With this in mind, we compared the genome of KNG with those of 10 goat populations in order to uncover selection signatures of KNG. The 10 populations were *C. aegagrus*, ING, MNG, KCB, AN, BA, KS, KB, the entire Saanen group (AS, KS, SS and FS), and the entire Boer group (AB, KB and SB). To consider the overall genomic characteristics, the sub-populations of Boer and Saanen were pooled as the entire Boer and Saanen groups, respectively. We then searched for extended linked regions with extreme haplotype homozygosity and highly differentiated regions with variations of allele frequency, using cross-population extended haplotype homozygosity (XP-EHH) (Sabeti et al., 2007) and cross-population composite likelihood ratio (XP-CLR) (Chen et al., 2010) analyses. The XP-EHH method, based on the extended haplotype homozygosity concept, is not sensitive to allele frequencies and is effective for the unreliable demographic model (Sabeti et al., 2007). The XP-CLR method, based on the composite likelihood ratio test, has the advantage of effectively detecting selective sweep regions when a population has a simple structure, a low migration rate, or difficulty in estimating the local recombination rate (Racimo, 2016). Therefore, these methods were appropriate for our study. Particularly, the approach combining these two methods has been reported to be able to increase the power to pinpoint selected regions, and has been used widely to uncover genes involved in local adaptations (Vatsiou et al., 2016). After analysis, we set a strict cut-off line in order to exclude false-positive results due to the genetic drift as many as possible. We considered outlier regions belonging to the top 0.1% of the empirical distributions of XP-EHH and XP-CLR statistics as candidate regions (**Data sheet 1: Figures S9** and **S10**, and **Data sheet 3** and **4**). Genes corresponding to these regions were annotated as candidate selected genes (**Data sheet 1: Figures S11A–J** and **Table S10**). The candidate genes derived from these two methods were pooled into one gene set for each population, in order to consider all genes that had undergone recent, soft, or hard sweeps. Then, gene set enrichment analysis (GSEA) (Stöckel et al., 2016) was performed for each gene set of each population to search for evidence of adaptation processes due to environmental selection. As a result, we found that KNG has selection signals for *Salmonella* infection pathway and cardiomyopathy pathway, respectively (**Table 1** and **Data sheet 5**).

Selection leaves detectable patterns in linkage disequilibrium, genetic diversity, and site frequency spectrum at the genome level, since it modifies the neutral pattern of the genomic region under the neutral theory of molecular evolution (Ross-Ibarra et al., 2007; Qanbari et al., 2010). When an allele frequency of a specific locus is affected by the selection, allele frequencies of closely linked loci around the locus are also affected, unlike the random process of genetic drift (Nielsen et al., 2005; Gianola et al., 2010; Qanbari et al., 2011). Therefore, we further investigated

patterns of nucleotide diversity, linkage disequilibrium, haplotype diversity, and Fst of LBP, BPI, ITGB6, and TTN genes, among KNG's candidate genes enriched in *Salmonella* infection and cardiomyopathy pathways (**Table 2**, **Figures 4A–D**, **Figures 5A–D**, and **Data sheet 1: Figures S12–S16**). These genes revealed traces of the environmental selection with the genetic drift that KNG underwent.

The Adaptive Characteristics of KNG to *Salmonella* Infection

In previous experimental studies, Jang and Kang reported that *Salmonella* species were not isolated from the feces of 49 KNG with diarrhea symptoms or 620 healthy KNG (Jang, 1995; Kang and Tak, 1996). Lee suggested that the reason for the absence of *Salmonella* infection in KNG was due to their excellent antibody productivity and inherent resistance factors (Lee et al., 2000). Through the selection analysis and the GSEA comparing KNG with 10 goat populations, we found that KNG showed selection signals for the *Salmonella* infection pathway in nine goat populations, excluding KCB (**Table 1** and **Data sheet 5**). The KCB, which was formed by hybridizing KNG with other goat breeds, is presumed to preserve a substantial portion of the genomic characteristics derived from KNG.

We further confirmed the LBP (Lipopolysaccharide Binding Protein) and BPI (Bactericidal Permeability Increasing Protein) genes, among 11 genes showing selection signals in the *Salmonella* infection pathway (**Table 2** and **Figures 4A**, **B**). The LPB gene, which encodes a lipopolysaccharide-binding protein, binds to the lipid A moiety of bacterial lipopolysaccharides to promote the release of cytokines, and the BPI gene, which encodes the bactericidal/permeability-increasing protein, regulates the LPS-dependent monocyte responses by binding to LPS along with the product of the LBP gene. The LBP gene plays an important role in the innate immune response of organisms (Wilde et al., 1994; Eckert et al., 2013), and the BPI gene plays an important role in antimicrobial activity against gram-negative organisms, as a paralogue of the LBP gene (Brister et al., 2014). Throughout the entire region of each gene, KNG showed the low nucleotide diversity and haplotype diversity patterns and presented a distinctive haplotype sharing pattern (a yellow square in **Figures 4A**, **B**, and **Data sheet 1: Figures S12A–D**). The pattern of the almost pure haplotype homozygosity, which distinguished noticeably from other populations, provides evidence that KNG has experienced the strong genetic drift which extensively affected the frequency of the alleles. Additionally, we discovered one selective sweep region with one missense variant in each gene, where KNG has been affected by their environment (**Figures 4C**, **D**). One missense variant was found in the 68,750,237 bp position (p.Asp217Glu) of the LBP gene (**Figure 4C**), and the other was found in the 68,695,875 bp position (p.Gln104Arg) of the BPI gene (**Figure 4D**). The haplotype frequencies of these variants in the LBP and BPI genes were the lowest in KNG, at 0.052 and 0.017, respectively, when excluding FA, SA, FS, and SB, which have low sample sizes. Conversely, the haplotype frequencies without these variants were the highest in KNG, at 0.930 and 0.983, respectively.

We confirmed that these LBP and BPI genes have been fixed or are being fixed in the direction of conserving their function in KNG. These results suggest that KNG has been more stabilized than other breeds for antimicrobial activity against gram-negative organisms as well as the innate immune response to *Salmonella* infection. We propose that KNG has accumulated local environmental pressure along with gene drift and has partially adapted to the *Salmonella* infection. In **Figures 4C**, **D**, only the top two haplotype frequencies for each goat population are illustrated, due to the limitations of illustration size. All haplotype frequencies are provided in **Data sheet 6**.

TABLE 2 | Candidate genes showing distinct patterns among genes involved in *Salmonella* infection and cardiomyopathy terms of KNG (See **Data sheet 3** and **4** for summary of all selected genes).

Selected genes	Association	CHR[a]	XP-CLR	XP-EHH		Candidate SNP position	Selected breeds[d]
			Score range[b]	Score range[b]	*p*-value range[c]		
LBP	*Salmonella* infection	13	(9.27, 9.27)	(3.31, 4.52)	(2.91, 3.33)	68,750,237 (p.Asp217Glu)	CA, IN, MN, KCB, TB, KB
BPI	Antimicrobial activity	13	(6.74, 9.96)	(3.31, 4.87)	(2.91, 3.33)	68,695,875 (p.Gln104Arg)	CA, IN, MN, KCB, TB, TS, BA
ITGB6	Cardiomyopathy	2	(8.11, 11.57)	(3.4, 3.4)	(3.16, 3.16)	–	CA, IN, MN, KCB, TB, KS, AN, BA
TTN	Cardiomyopathy	2	(7.94, 13.81)	(3.46, 3.72)	(2.9, 3.45)	19,127,870 (p.Ile1202Thr) 19,167,388 (p.Ala3702Thr) 19,188,702 (p.Val7638Ile)	CA, IN, MN, KCB, TB, TS, KS

Dash (-) indicates non-significant results

[a]*Chromosome*

[b]*The range of minimum and maximum values of XP-CLR and XP-EHH scores which each selected breed has for the selected gene.*

[c]*The range of minimum and maximum values of - log p-values for the XP-EHH scores, derived from the empirical distribution.*

[d]*Abbreviations in the selected breeds column means: IN is Iranian native goat, MN is Moroccan native goat, CA is C. aegagrus, KCB is Korean crossbred, KB is Korean Boer, KS is Korean Saanen, BA is British Alpine, AN is Anglo-Nubian, TB is the entire Boer group, and TS is the entire Saanen group.*

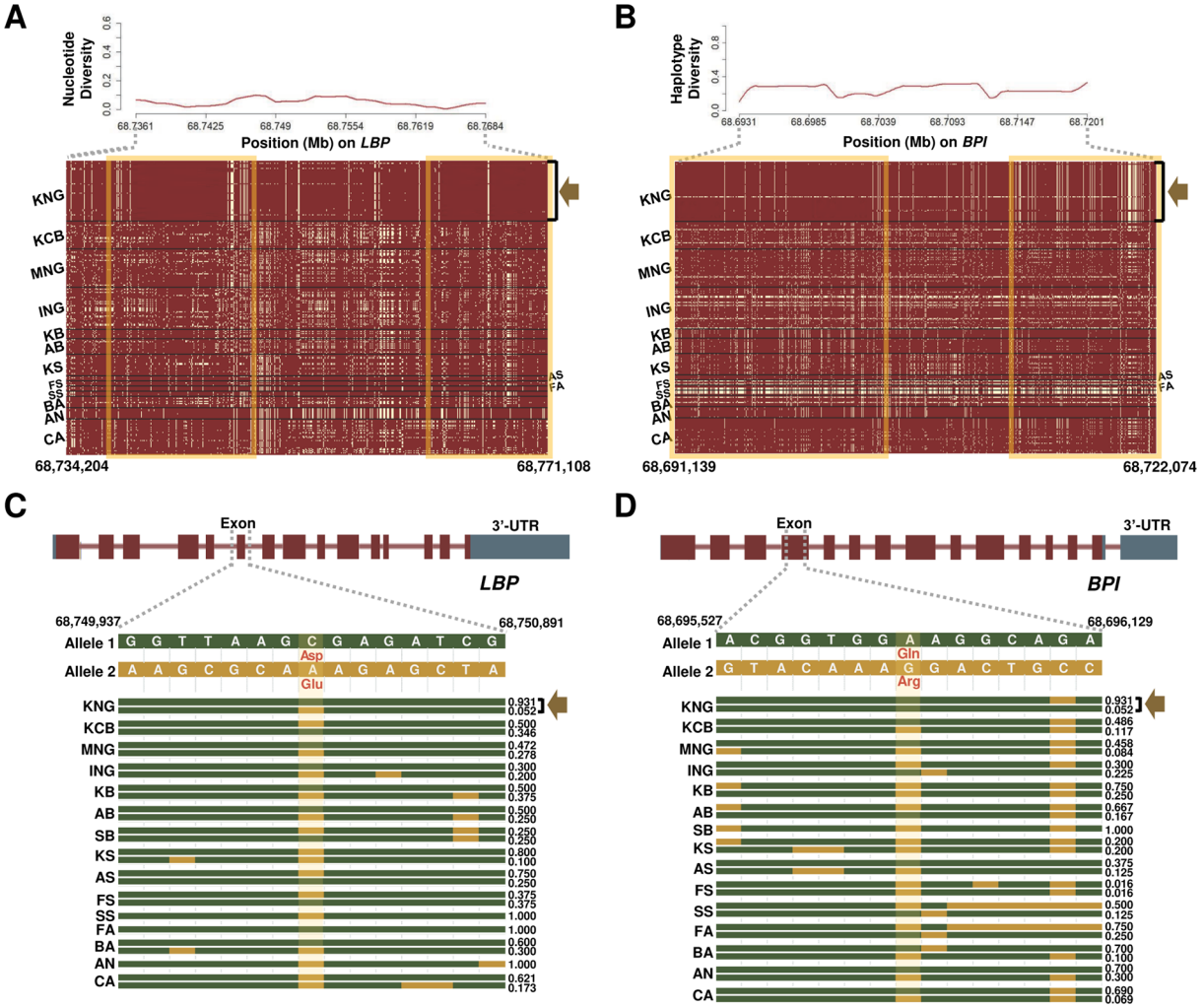


FIGURE 4 | Selection signature for *Salmonella* infection in KNG. **(A)** Nucleotide diversity (above) and haplotype sharing (bottom) patterns for the region of 68,734,204-68,771,108-bp of the LBP gene located on chromosome 13. **(B)** Haplotype diversity (above) and haplotype sharing (bottom) patterns for the region of 68,691,139-68,722,074-bp of the BPI gene located on chromosome 13. In the haplotype sharing plots, the yellow rectangle highlights the pattern in which KNG is differentiated from other goat populations. **(C–D)** Gene structures and haplotype frequencies of regions containing a missense SNP in LBP and BPI genes, respectively. The missense SNPs, highlighted in yellow, represent p.Asp217Glu on the 68,750,237 bp position (LBP) and p.Gln104Arg on the 68,695,875 bp position (BPI) (See **Data sheet 6**).

The Adaptive Characteristics of KNG to Cardiomyopathy Challenge

Cardiomyopathy is any disease affecting the muscle, size, and shape of the heart, and is presents as three main types: dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and arrhythmogenic right ventricular cardiomyopathy (ARVC) (Rush et al., 2002). It is mostly idiopathic, and symptoms are different ranging from no symptoms, to difficulty breathing, to sudden death. This disease has been reported to occur not only in humans but also in dogs and cats (Broschk and Distl, 2005; Meurs et al., 2007), and even in the Saanen goat breed (Tontis et al., 1992). Using the selection analysis and the GSEA, we found that KNG has selection signals for all of DCM, HCM, and ARVC pathways in eight goat populations excepting for KCB and BA (**Table 1** and **Data sheet 5**). The candidate genes for KCB and BA were significantly enriched only in DCM pathway and ARVC pathway, respectively. Although further research is needed, KCB is presumed to preserve a significant portion of the genomic characteristics derived from KNG, and BA is presumed to have partially adapted to this disease due to the artificial or environmental effects (**Figure 3A**).

We further confirmed the ITGB6 (Integrin Subunit Beta 6) and TTN (Titin) genes among the several genes showing selection signals in the three cardiomyopathy pathways (**Table 2** and **Figures 5A**, **B**). The ITGB6 gene, which encodes a protein of the integrin superfamily, is involved in all of the DCM, HCM, and ARVC pathways, and it has been reported to be particularly closely related to the ARVC pathway (O'Leary et al., 2015; Stelzer et al., 2016). The TTN gene, which encodes a large abundant protein of striated muscle containing cardiac muscle tissues, is involved in the DCM and HCM pathways, and it has been reported as one of the positively selected genes that influence cardiomyopathy in a bear

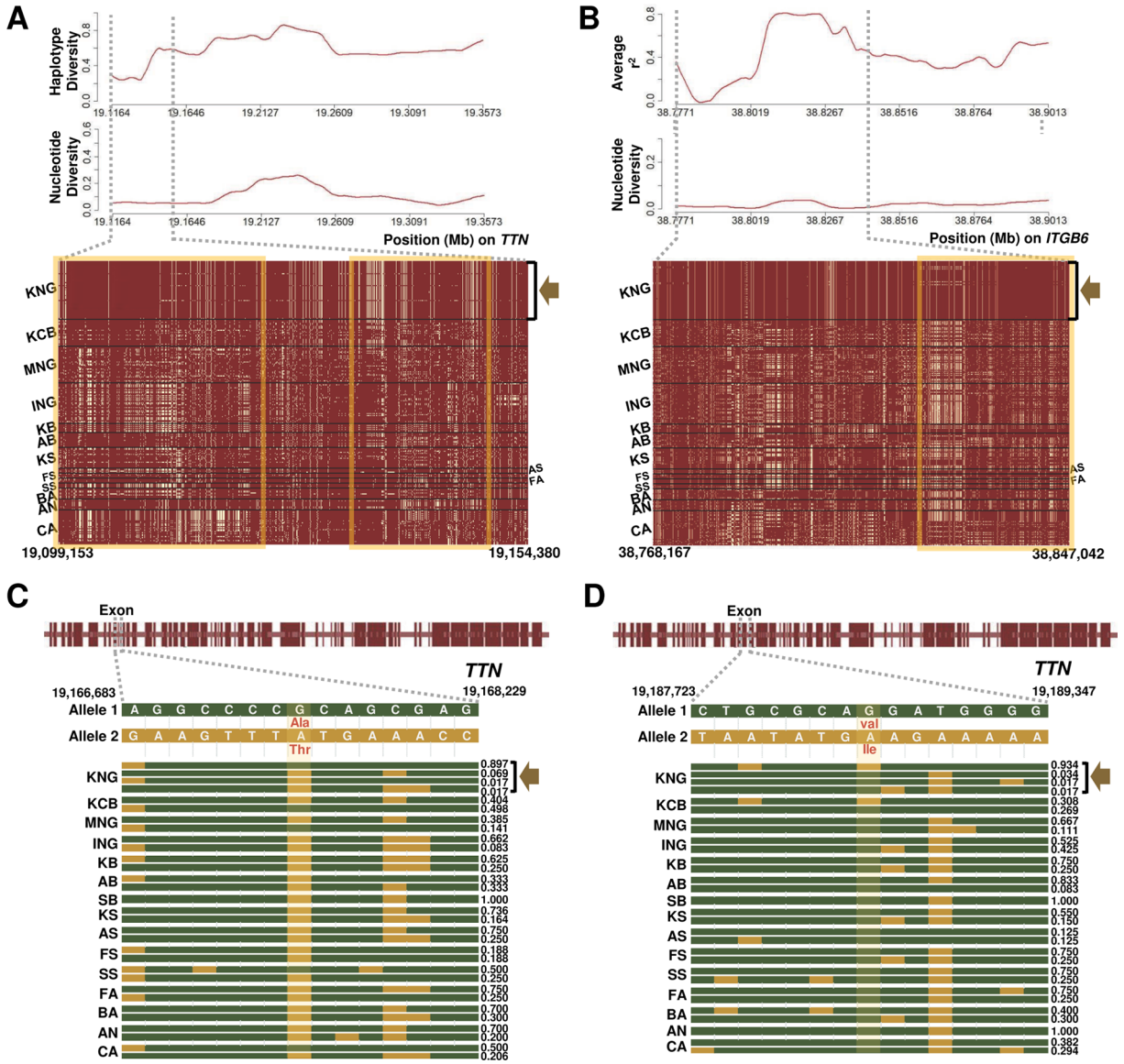


FIGURE 5 | Selection signature for cardiomyopathy in KNG. **(A)** Haplotype and nucleotide diversity patterns (above) and haplotype sharing pattern (bottom) for the 19,099,153-19,154,380-bp region of the TTN gene located on chromosome 2. **(B)** Average linkage disequilibrium and nucleotide diversity patterns (above) and haplotype sharing pattern (bottom) for the 38,768,167-38,847,042-bp region of the ITGB6 gene located on chromosome 2. In the haplotype sharing plots, the yellow rectangle highlights the pattern in which the KNG is differentiated from other goat populations. **(C–D)** Gene structures and haplotype frequencies of regions containing missense SNPs in the TTN gene. The missense SNPs, highlighted in yellow, represent p.Val7638Ile on the 19,188,702 bp position and p.Ala3702Thr on the 19,167,388 bp position, respectively (See **Data sheet 6**).

breed (Liu et al., 2014). In both genes, KNG showed the lowest nucleotide diversity and haplotype diversity patterns and presented an almost pure haplotype sharing pattern as a result of the genetic drift (A yellow square in **Figures 5A**, **B** and **Data sheet 1: Figures S13–S16**). In addition, KNG showed traces of selective sweeps with three missense variants in the TTN gene (**Figures 5C**, **D** and **Data sheet 1: Figure S16A**). We screened the regions containing these missense mutations along with haplotype frequencies. The haplotype frequency with a missense SNP (p.Ile1202Thr) found at the 19,127,870 bp position was highest in KNG at 0.948, followed by in KCB and KB at 0.538 and 0.375, respectively. (**Data sheet 1: Figure S16A**). This missense SNP showed a tendency to hitchhike the SNPs of 19,127,266 bp and 19,128,208 bp positions together. Another missense SNP (p.Ala3702Thr) found at the 19,167,388 bp position showed a tendency to replace the SNP of 19,167,677 bp position with the reference SNP (**Figure 5C**). The haplotype frequency of this region was the highest in KNG at 0.897, followed by in SS and KCB at 0.498 and 0.500, respectively. Most goat populations possessed this missense mutation, but KNG maintained this SNP as the reference variant with a high frequency. The other missense SNP was found at the 19,188,702 bp position (**Figure 5D**). This variant showed a tendency to replace the SNP

of 19,188,088 bp position with an alternative SNP and the SNP of 19,189,131 bp position with a reference SNP, respectively. The haplotype frequency was the highest in KNG at 0.934, followed by in KCB at 0.404. We further identified the region where the ITGB6 gene has been affected by the selective sweep. KNG showed the highest average LD with the lowest nucleotide diversity in the region of 38,805,100 bp–38,833,000 bp (**Figure 5B** and **Data sheet 1: Figures S13A–C** and **S14A–B**).

We confirmed that these ITGB6 and TTN genes have been affected by the local environment along with the genetic drift. Particularly, the coexistence of three missense SNPs with the highest and the lowest frequencies in KNG suggests that this TTN gene has been playing a functional role in adapting to cardiomyopathy as one of several candidate genes. Based on our genomic research, we propose that KNG has partially adapted to the cardiomyopathy under their various environmental pressure.

DISCUSSION

Genomic Characteristics of KNG

Domestication and subsequent geographical expansion have generated a variety of indigenous livestock breeds. These breeds have accumulated multiple genetic variations affecting a variety of traits over time and have developed their own unique genomic characteristics in the course of enhancing their fitness in different local environments. These genomic characteristics are important as a genomic basis for coping with future threats to the species arising from environmental change (Benjelloun et al., 2015), but are rapidly disappearing due to extensive crossbreeding and substitution with imported breeds. Therefore, to reveal their unique genomic characteristics, many genomic studies have been carried out in various indigenous livestock: cattle (Browett et al., 2018; Weldenegodguad et al., 2018); chicken (Johansson and Nelson, 2015; Walugembe et al., 2018); sheep (Yang et al., 2016; Edea et al., 2017); and goat (Benjelloun et al., 2015; Cao et al., 2019). In this context, our study focused on identifying KNG and revealed their distinct genomic characteristics.

To investigate KNG in detail, we utilized the whole-genome variations of a total of 10 goat breeds, including three indigenous breeds (KNG, ING, and MNG), six commercial breeds (Saanen breed, Boer breed, AN, BA, FA, and, KCB), and one ancestral species (*C. aegagrus*). A total of 38,658,962 bi-allelic SNPs were detected in 29 autosomes of 10 breeds, and we identified that these SNPs covered a considerable portion of their reference genome at an average distance of 64.88 bases between SNPs (**Data sheet 1: Table S4**). With the exception of their ancestral species, the number of bi-allelic SNPs was the highest for KNG (37,715,208) and followed by ING, KCB, and MNG (35,742,191, 33,464,841, and 32,914,220) (**Figure 1B** and **Data sheet 1: Table S5**). In respect of π and LD calculated using these bi-allelic SNPs, the KNG exhibited the lowest π (0.001472) and the highest LD (0.088431) among three indigenous breeds (**Data sheet 1: Table S7**). Particularly, the KNG's π value was consistent with the adjusted π value reported by a previous study (calculation window size adjusted from 1Mb to 100Kb) (Lee et al., 2016). Considering their low π and many SNPs, our results indicate that the KNG has a fair number of homozygous SNP variants distinguished from other goat breeds, relatively. In addition to this, the high LD value implies that their homozygous SNP variants have a high level of association with each other due to evolutionary pressures such as selection or genetic drift.

The population analyses conducted through various methodologies supported our hypothesis that the KNG has unique genomic characteristics, which are distinct from those of other goat breeds. Within a large category of genomic diversity parameters, the genomic features of the eight goat breeds did not show large differences, but KNG and *C. aegagrus* showed distinctive genomic characteristics (**Figures 2A–D**). The KNG was separated to the rightmost in PCA, showed a near-identical genomic composition in structure analysis (**Figure 2B**), and exhibited high levels of genetic differentiation compared with other goat breeds (**Data sheet 1: Table S8**). Our results additionally confirmed that the genomic composition of KNG (blue color) coincided with one of ING, and another genomic composition of ING (green color) was consistent with one of their ancestral species (**Figure 2B**). Particularly, the ING inhabiting the region of Iran where *C. hircus* was first domesticated showed a linear relationship with the *C. aegagrus*, and they positioned at the center of the 10 goat breeds in the PCA (**Figure 2C**). Given the origins of *C. hircus* and KNG, these results suggest that the ING has maintained a substantial portion of genomic characteristics derived from its ancestral species since the domestication, and that the KNG has formed its own genomic characteristics since influx into the Korean Peninsula about 2,000 years ago (Tavakolian, 2000; Zeder and Hesse, 2000). Meanwhile, a previous study reported that the ING inhabiting the north of the Zagros mountain has the most similar genomic structure to their ancestor, *C. aegagrus* (Vahidi et al., 2014). In our study, the genomic compositions of ING samples were almost identical to those of ING samples which have been reported to be the indigenous goats of the north Zagros mountain. This result indicates that the ING samples were suitable for our study to compare KNG with various goat breeds, as the closest domesticated goats to their ancestral species.

From the analyses of the gene flow and N_e, we revealed that the KNG's unique genomic characteristics are associated, at least in part, with their isolated environment (**Figures 3A**, **B**). We confirmed that the KNG underwent a severe genetic bottleneck event as they entered the Korean Peninsula about 2,000 years ago (**Figure 3B**), and have experienced little genetic interactions with other breeds (only except for KCB) (**Figure 3A**). To clarify the interaction signals of KNG, the D-statistic (Durand et al., 2011) and 3-population (Reich et al., 2009) tests were also performed, but no signal was detected except for the KCB (**Data sheet 2**). These results indicate that the KNG has accumulated their local environmental pressure for a long time, and has developed their own genomic characteristics with little genetic interaction with other breeds. Also, as genomic evidence, these results support the previous studies which reported on the origin and isolated environment of KNG (Son, 1999; Tavakolian, 2000; Odahara et al., 2006). So far, we revealed the unique genomic characteristics of KNG through a comparison of 10 goat breeds. We expect that our detailed review for the KNG including other goat breeds would contribute to the establishment of biodiversity conservation strategies regarding indigenous goats.

Adaptive Characteristics of KNG

During long-term adaptation to the various environments, indigenous livestock breeds have developed their own adaptive characteristics which enhance fitness to harsh environments or resistance to specific diseases. These characteristics have provided an important genetic basis for various breeding programs to improve livestock (Guan et al., 2016). Thus, to identify their adaptive characteristics, many studies on selection signatures have been conducted in various indigenous livestock: cattle (Taye et al., 2017); chicken (Lawal et al., 2018); goat (Guo et al., 2018); sheep (Liu et al., 2016); and pig (Li et al., 2014). From this perspective, our study compared KNG with other 10 goat breeds, and revealed that the KNG has selection signatures for *Salmonella* infection and cardiomyopathy pathways (**Table 1**).

Salmonella infection has effects ranging from growth delay to livestock death (Cummings et al., 2010). The identification of indigenous breeds adapted to this infection can be valuable in livestock breeding programs for enhancing the survival rate and preventing disease transmission. However, there have been few investigations into indigenous livestock breeds which carry this resistance, except for the Sri Lankan indigenous chicken (Weerasooriya et al., 2017) and the KNG (Jang, 1995; Kang and Tak, 1996; Lee et al., 2000). Although these two breeds have been reported to be resistant to *Salmonella* infection through experimental studies, their utilization in breeding programs has been limited due to the lack of genomic studies. In this study, we identified that the KNG exhibits selection signals with respect to the *Salmonella* infection pathway for nine goat breeds except for KCB (**Table 1**). To clarify the KNG's selection signals, we further examined the LBP and BPI genes among their 11 candidate genes (**Table 2**). The KNG showed low nucleotide and haplotype diversity patterns and a unique haplotype-sharing pattern over the entire region of these genes (**Figures 4A**, **B**). Also, as a consequence of strong selection pressures, the KNG exhibited selective sweep regions with one missense SNP variant in each gene (**Figures 4C**, **D**). The haplotype frequencies containing these missense variants were the lowest in KNG when excluding FA, SA, FS, and SB with low sample sizes (**Data sheet 6**). Considering the functions of two genes, these results indicate that KNG has been more stabilized than other breeds for the antimicrobial activity to gram-negative organisms and the innate immune response to *Salmonella* infection. Our results provide genomic evidence to support previous biological studies, and statistically, propose that KNG has adaptive characteristics for *Salmonella* infection.

As one of the novel adaptive characteristics, we observed that the KNG has selection signals for all three types of cardiomyopathy pathways in eight goat breeds (**Table 1**). The exceptions were KCB and BA, and the KNG's candidate genes were significantly enriched only in DCM pathway for KCB and only in ARVC pathway for BA. Among the KNG's candidate genes, we further investigated the TTN (associated with DCM and HCM) and ITGB6 (associated with ARVC) genes which show distinctive selection patterns (**Table 2** and **Figures 5A**, **B**). For TTN, the KNG exhibited a trace of selective sweep together with hitchhiking effects in three missense SNP variants (**Figures 5C**, **D**, and **Data sheet 1: Figure S16A**). In the 38,805,100–38,833,000 bp region of the ITGB6 gene, the KNG showed the highest LD and almost pure haplotype patterns due to strong selection pressure (**Figure 5B**, and **Data sheet 1: Figure S14A-B**). These results indicate that the ITGB6 and TTN genes, particularly, have been playing a functional role in adapting to cardiomyopathy in the KNG. Based on our genomic research, we statistically propose that the KNG has partially adapted to cardiomyopathy.

In our results, the KNG did not show selection signals for *Salmonella* infection in KCB and for cardiomyopathy in KCB (HCM and ARVC) and BA (DCM and HCM). The KCB was recently formed by hybridization of KNG with other goat breeds, in order to improve various traits of the KNG (Lee et al., 2016). The KCB shared a large amount of genomic composition with KNG in structure analysis (**Figure 2B**) and exhibited a similar genomic characteristic to KNG in PCA (**Figures 2C**, **D**). Also, they showed substantial interactions with the KNG in gene flow analysis, Patterson's D-statistic test, and the 3-Population test (**Figure 3A** and **Data sheet 2**). These results indicate that the KCB has acquired a considerable portion of their genome characteristics from KNG, and that the purpose of the crossbreeding program has been achieved to a large extent. However, considering their still high π (0.001908) and low LD (0.068539) (**Data sheet 1: Table S7**), it is suggested that the KCB need an additional breeding program to stabilize their genomic and adaptive characteristics. Meanwhile, the BA was developed in the Swiss and Austrian Alps in the early 1900s and introduced into Australia in about 1960. Our BA samples collected in Australia showed the lowest π (0.001251) and the highest LD (0.288801), except for the AN (π: 0.001117, and LD: 0.327566) (**Data sheet 1: Table S7**). These genomic characteristics imply that the BA had undergone significant genetic drift upon being introduced to Australia and have experienced multiple selection events. Our study confirmed the possibility that the BA may have partially adapted to cardiomyopathy in their environment, but we propose further research to clarify this adaptive characteristic, due to their small sample size.

Our study has several limitations. First, the SNP variants of some breeds (FA, AS, SS, FS, and SB) may have been affected by SNP ascertainment bias due to their small sample sizes. Their SNP variants may not have adequately represented their entire breeds, and some analysis results for them may have been distorted. Therefore, to minimize this problem, our study utilized these breeds as only references against which to compare the genomic characteristics of the other breeds. In contrast, we could avoid another SNP ascertainment bias due to SNP discovery protocols by using the whole-genome sequencing protocol. In the case of using Illumina's Goat SNP50 BeadChip (Tosser-Klopp et al., 2014), which contains approximately 53,346 SNP variants, some results of the population analyses could be distorted due to this bias, since its SNP markers cover neither all goat breeds nor entire genomic regions (Lachance and Tishkoff, 2013). Second, some adaptive characteristics for KNG have been identified, but have not been validated by biological experiments. To minimize this limitation, we conducted rigorous statistical calculations. We compared the KNG with other 10 goat breeds by using two selection analysis methods, XP-CLR and XP-EHH, and detected candidate selected genes using strict cut-offs. We then confirmed the KNG's adaptation signals

through the GSEA of the candidate genes, and revealed the genomic regions affected by the selection pressure in some candidate genes. Despite these efforts, our results still require further experimental validation, but we anticipate that these candidate genes and their targeted genomic regions will be helpful in future experimental studies aimed at identifying the characteristics of KNG.

Although our study has some limitations, our catalog of genome characteristics of 10 goat breeds would provide the basis for establishing various appropriate breeding strategies. Also, our findings on the genomic and adaptive characteristics of the KNG will help to set directions of biodiversity conservation programs as well as crossbreeding and grading-up programs for improving goat breeds.

CONCLUSION

The valuable genomic characteristics that indigenous breeds have accumulated for a long time are being threatened by crossbreeding with imported breeds with high productivity. Particularly in the case of Korea, KNG is rapidly being substituted with KCB, which was formed by hybridizing KNG with other breeds to improve KNG's inferior commercial traits. Although their characteristics may not be commercially valuable, they could have unique abilities to survive in a particular environment or disease. In this respect, our research on the genomic population dynamics of KNG, including various goat breeds, provides an important basis for establishing a direction for biodiversity conservation strategies. Although our findings for adaptive characteristics have a limitation that is provided without biological validation, these are expected to not only provide new and other options to those seeking to improve the viability and the resilience of goats but also present targeted genomic regions to *in vivo* or *in vitro* studies trying to employ our hypothesis. In addition, our newly generated whole-genome data that is opened to the public database will contribute to the knowledge for further research.

ETHICS STATEMENT

This study was carried out in accordance with the guidelines of the Institutional Animal Care and Use Committee (IACUC) and was approved by the National Institute of Animal Science, Rural Development Administration, Republic of Korea (Approval No: 2012-D-010). The animal preparation and experimentation were conducted in accordance with the protocol approved by the guidance of the IACUC.

AUTHOR CONTRIBUTIONS

NK conceived and supervised this project. J-YK, and SJ performed the data analysis and wrote the draft manuscript. KHK, and W-JL supported the data analysis and interpretation. H-YL assisted with the literature search and figure preparation. All authors discussed the results and read and approved the final manuscript.

FUNDING

This project was supported by grants from the National Research Foundation of Korea (NRF-2014M3C9A3064552), the KRIBB Initiative program, and the Cooperative Research Program for Agriculture Science and Technology Development Project of Rural Development Administration (Republic of Korea) (No. PJ00868002).

ACKNOWLEDGMENTS

We deeply thank the collaborators of National Institute of Animal Science, which kindly provided a huge amount of data as part of the Initiative of Animal Genome Open Data Utilization Plan held on March 10, 2016.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00699/full#supplementary-material

DATA SHEET 1 | Supplemental information including Note, Tables S1-S10 and Figures S1-S16 (DOC 8,082 kb).

DATA SHEET 2 | The results of D-statistic and three-population tests (XLSX 191 kb).

DATA SHEET 3 | The results of XP-EHH analysis (XLSX 483 kb).

DATA SHEET 4 | The results of XP-CLR analysis (XLSX 255 kb).

DATA SHEET 5 | The summary of gene set enrichment analysis (XLSX 39 kb).

DATA SHEET 6 | The summary of six non-synonymous variants and their surrounding haplotype frequencies (XLSX 49 kb).

REFERENCES

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Babayan, N. (2016). Goat Industry development project in ARMENIA, in *Sustainable Goat Breeding and Goat Farming in the Central and Eastern European Countries*. Ed. K. Sándor (Rome, FL: FAO), 51–56.

Ballou, J. D., Lees, C., Faust, L. J., Long, S., Lynch, C., Bingaman Lackey, L. et al., (2010). Demographic and genetic management of captive populations, in *Wild mammals in captivity: principles and techniques for zoo management*. Eds. D. G. Kleiman, K. V. Thompson, and C. K. Baer (Chicago, FL: The University of Chicago Press), 219–252.

Benjelloun, B., Alberto, F. J., Streeter, I., Boyer, F., Coissac, E., Stucki, S., et al. (2015). Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Front. Genet.* 6, 107. doi: 10.3389/fgene.2015.00107

Boitard, S., Rodríguez, W., Jay, F., Mona, S., and Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data-an approximate Bayesian computation approach. *PLoS Genet.* 12 (3), e1005877. doi: 10.1371/journal.pgen.1005877

Brister, J. R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2014). NCBI viral genomes resource. *Nucleic Acids Res.* 43 (Database issue), D571–D577. doi: 10.1093/nar/gku1207

Broschk, C., and Distl, O. (2005). Dilated cardiomyopathy (DCM) in dogs–pathological, clinical, diagnosis and genetic aspects. *Dtsch. Tierarztl. Wochenschr.* 112 (10), 380–385.

Browett, S., McHugo, G., Richardson, I. W., Magee, D. A., Park, S. D., Fahey, A. G., et al. (2018). Genomic characterisation of the indigenous Irish Kerry cattle breed. *Front. Genet.* 9, 51. doi: 10.3389/fgene.2018.00051

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 (5), 1084–1097. doi: 10.1086/521987

Cao, Y., Xu, H., Li, R., Gao, S., Chen, N., Luo, J., et al. (2019). Genetic basis of phenotypic differences between Chinese Yunling black goats and Nubian goats revealed by allele-specific expression in their F1 hybrids. *Front. Genet.* 10, 145. doi: 10.3389/fgene.2019.00145

Chakraborty, R., and Deka, R. (2005). Isolated populations, in *Encyclopedia of Biostatistics*, 2nd ed. Eds. P. Armitage and T. Colton (New York, FL: Wiley & Sons), 4. doi: 10.1002/0470011815.b2a05056

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20 (3), 393–402. doi: 10.1101/gr.100545.109

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6 (2), 80–92. doi: 10.4161/fly.19695

Cummings, K. J., Warnick, L. D., Elton, M., Gröhn, Y. T., McDonough, P. L., and Siler, J. D. (2010). The effect of clinical outbreaks of salmonellosis on the prevalence of fecal Salmonella shedding among dairy cattle in New York. *Foodborne Pathog. Dis.* 7 (7), 815–823. doi: 10.1089/fpd.2009.0481

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 (15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., et al. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat. Biotechnol.* 31 (2), 135–141. doi: 10.1038/nbt.2478

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28 (8), 2239–2252. doi: 10.1093/molbev/msr048

Eckert, J. K., Kim, Y. J., Kim, J. I., Gürtler, K., Oh, D.-Y., Sur, S., et al. (2013). The crystal structure of lipopolysaccharide binding protein reveals the location of a frequent mutation that impairs innate immunity. *Immunity* 39 (4), 647–660. doi: 10.1016/j.immuni.2013.09.005

Edea, Z., Dessie, T., Dadi, H., Do, K. T., and Kim, K. S. (2017). Genetic diversity and population structure of Ethiopian sheep populations revealed by high-density SNP markers. *Front. Genet.* 8, 218. doi: 10.3389/fgene.2017.00218

Ewens, W. (1990). The minimum viable population size as a genetic and a demographic concept, in *Convergent issues in genetics and demography*. Eds. J. Adams, D. A. Lam, A. I. Hermalin, and P. E. Smouse (New York, FL: Oxford University Press), 307–316.

Frankham, R., Ballou, J. D., and Briscoe, D. A., (2010). *Introduction to conservation genetics*. 2nd ed. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809002

Futuyma, D. J. (2009). Natural selection and adaptation, in *Evolution*. Ed. D. J. Futuyma (Massachusetts, FL: SINAUER ASSOCIATES), 279–301.

Gasca-Pineda, J., Cassaigne, I., Alonso, R. A., and Eguiarte, L. E. (2013). Effective population size, genetic variation, and their relevance for conservation: the bighorn sheep in Tiburon Island and comparisons with managed artiodactyls. *PLoS One* 8, e78120. doi: 10.1371/journal.pone.0078120

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14 (7), 685–695. doi: 10.1093/oxfordjournals.molbev.a025808

Gianola, D., Simianer, H., and Qanbari, S. (2010). A two-step method for detecting selection signatures using genetic markers. *Genet. Res.* 92 (2), 141–155. doi: 10.1017/S0016672310000121

Giovambattista, G., Ripoli, M. V., Peral-Garcia, P., and Bouzat, J. L. (2001). Indigenous domestic breeds as reservoirs of genetic diversity: the Argentinean Creole cattle. *Anim. Genet.* 32 (5), 240–247. doi: 10.1046/j.1365-2052.2001.00774.x

Guan, D., Luo, N., Tan, X., Zhao, Z., Huang, Y., Na, R., et al. (2016). Scanning of selection signature provides a glimpse into important economic traits in goats (Capra hircus). *Sci. Rep.* 6, 36372. doi: 10.1038/srep36372

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 (1), 10405. doi: 10.1038/s41598-018-28719-w

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38 (6), 226–231. doi: 10.1007/BF01245622

Jang, I. (1995). *Studies on bacterial and parasitic diseases of Korean native goats.* Suwon, Korea: Rural Development Administration.

Johansson, A. M., and Nelson, R. M. (2015). Characterization of genetic diversity and gene mapping in two Swedish local chicken breeds. *Front. Genet.* 6, 44. doi: 10.3389/fgene.2015.00044

Kang, M. (1967). Studies on the origin of Korean native goat. *Korean J. Anim. Sci.* 9, 1005–1010.

Kang, M., and Tak, R. (1996). Resistance of Korean native goats to Salmonella typhimurium. *Korean J. Vet. Public Health (Korea Republic)* 20, 27–36.

Kim, J.-H., Cho, C.-Y., Choi, S.-B., Cho, Y.-M., Yeon, S.-H., and Yang, B.-S. (2011). mtDNA diversity and phylogenetic analysis of Korean native goats. *J. Life Sci.* 21, 1329–1335. doi: 10.5352/JLS.2011.21.9.1329

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16 (2), 111–120. doi: 10.1007/BF01731581

Lachance, J., and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35 (9), 780–786. doi: 10.1002/bies.201300014

Lawal, R. A., Al-Atiyat, R. M., Aljumaah, R. S., Silva, P., Mwacharo, J. M., and Hanotte, O. (2018). Whole-genome resequencing of red junglefowl and indigenous village chicken reveal new insights on the genome dynamics of the species. *Front. Genet.* 9, 264. doi: 10.3389/fgene.2018.00264

Lee, C., Lee, C., and Kwag, H. (2000). Studies on the diseases of the korean native goat-a review. *J. Vet. Clin.* 17, 32–44.

Lee, W., Ahn, S., Taye, M., Sung, S., Lee, H.-J., Cho, S., et al. (2016). Detecting positive selection of korean native goat populations using next-generation sequencing. *Mol. Cells* 39 (12), 862. doi: 10.14348/molcells.2016.0219

Li, J. Y., Chen, H., Lan, X. Y., Kong, X. J., and Min, L. J. (2008). Genetic diversity of five Chinese goat breeds assessed by microsatellite markers. *Czech J. Anim. Sci.* 53 (8), 315–319. doi: 10.17221/347-CJAS

Li, J., and Zhang, Y. (2009). Advances in research of the origin and domestication of domestic animals. *Biodivers. Sci.* 17 (4), 319–329. doi: 10.3724/SP.J.1003.2009.09150

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics* 26 (5), 589–595. doi: 10.1093/bioinformatics/btp698

Li, M., Tian, S., Yeung, C. K., Meng, X., Tang, Q., Niu, L., et al. (2014). Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Sci. Rep.* 4, 4678. doi: 10.1038/srep04678

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* 157 (4), 785–794. doi: 10.1016/j.cell.2014.03.054

Liu, Z., Ji, Z., Wang, G., Chao, T., Hou, L., and Wang, J. (2016). Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. *BMC Genomics* 17 (1), 863. doi: 10.1186/s12864-016-3212-2

Losos, J. B., and Ricklefs, R. E. (2009). Adaptation and diversification on islands. *Nature* 457, 830. doi: 10.1038/nature07893

Meurs, K. M., Fox, P. R., Norgard, M., Spier, A. W., Lamb, A., Koplitz, S. L., et al. (2007). A prospective genetic evaluation of familial dilated cardiomyopathy in the Doberman pinscher. *J. Vet. Intern. Med.* 21 (5), 1016–1020. doi: 10.1111/j.1939-1676.2007.tb03058.x

Naderi, S., Rezaei, H.-R., Pompanon, F., Blum, M. G., Negrini, R., Naghash, H.-R., et al. (2008). The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci. U. S. A.* 105 (46), 17659–17664. doi: 10.1073/pnas.0804782105

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15 (11), 1566–1575. doi: 10.1101/gr.4252305

Nomura, K., Yonezawa, T., Mano, S., Kawakami, S., Shedlock, A. M., Hasegawa, M., et al. (2013). Domestication process of the goat revealed by an analysis of the nearly complete mitochondrial protein-encoding genes. *PLoS One* 8 (8), e67775. doi: 10.1371/journal.pone.0067775

Odahara, S., Chung, H., Choi, S., Yu, S., Sasazaki, S., Mannen, H., et al. (2006). Mitochondrial DNA diversity of Korean native goats. Asian-Australas. *J. Anim. Sci.* 19 (4), 482–485. doi: 10.5713/ajas.2006.482

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2015). Reference sequence (RefSeq) database at NCBI: current status,

taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44 (D1), D733–D745. doi: 10.1093/nar/gkv1189

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

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7 (2), e30619. doi: 10.1371/journal.pone.0030619

Pergams, O. R., and Lawler, J. J. (2009). Recent and widespread rapid morphological change in rodents. *PLoS One* 4 (7), e6452. doi: 10.1371/journal.pone.0006452

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8 (11), e1002967. doi: 10.1371/journal.pgen.1002967

Porter, V., Alderson, L., Hall, S. J. G., and Sponenberg, D. P., (2016). *Mason's world encyclopedia of livestock breeds and breeding*. Wallingford: CABI Publishing. doi: 10.1079/9781845934668.0000

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 (3), 559–575. doi: 10.1086/w519795

Qanbari, S., Pimentel, E., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A., et al. (2010). A genome-wide scan for signatures of recent selection in Holstein cattle. *Anim. Genet.* 41 (4), 377–389. doi: 10.1111/j.1365-2052.2009.02016.x

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., et al. (2011). Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* 12, 318. doi: 10.1186/1471-2164-12-318

Racimo, F. (2016). Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* 202 (2), 733–750. doi: 10.1534/genetics.115.178095

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197 (2), 573–589. doi: 10.1534/genetics.114.164350

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature* 461 (7263), 489–494. doi: 10.1038/nature08365

Rischkowsky, B., and Pilling, D. (2007). *The state of the world's animal genetic resources for food and agriculture*. Rome: FAO.

Ross-Ibarra, J., Morrell, P. L., and Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8641–8648. doi: 10.1073/pnas.0700643104

Rush, J. E., Freeman, L. M., Fenollosa, N. K., and Brown, D. J. (2002). Population and survival characteristics of cats with hypertrophic cardiomyopathy: 260 cases (1990–1999). *J. Am. Vet. Med. Assoc.* 220 (2), 202–207. doi: 10.2460/javma.2002.220.202

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449 (7164), 913–918. doi: 10.1038/nature06250

Sodiq, A. (2004). *Doe productivity of kacang and peranakan etawah goats and factors affecting them in Indonesia. Dissertation/doctoral thesis*. Kassel (IL): University of Kassel.

Son, Y. S. (1999). Production and uses of korean native black goat. *Small Rumin. Res.* 34 (3), 303–308. doi: 10.1016/S0921-4488(99)00081-4

Song, S., Yao, N., Yang, M., Liu, X., Dong, K., Zhao, Q., et al. (2016). Exome sequencing reveals genetic differentiation due to high-altitude adaptation in the Tibetan cashmere goat (Capra hircus). *BMC Genomics* 17 (1), 122. doi: 10.1186/s12864-016-2449-0

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr. Protoc. *Bioinformatics* 54, 1.30.1–1.30.33. doi: 10.1002/cpbi.5

Stöckel, D., Kehl, T., Trampert, P., Schneider, L., Backes, C., Ludwig, N., et al. (2016). Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics* 32 (10), 1502–1508. doi: 10.1093/bioinformatics/btv770

Szpiech, Z. A., and Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31 (10), 2824–2827. doi: 10.1093/molbev/msu211

Taberlet, P., Valentini, A., Rezaei, H., Naderi, S., Pompanon, F., Negrini, R., et al. (2008). Are cattle, sheep, and goats endangered species? *Mol. Ecol.* 17 (1), 275–284. doi: 10.1111/j.1365-294X.2007.03475.x

Tavakolian, J. (2000). *An introduction to genetic resources of native farm animals in Iran*. Tehran (Iran): Animal Science Genetic Research Institute Press.

Taye, M., Lee, W., Caetano-Anolles, K., Dessie, T., Hanotte, O., Mwai, O. A., et al. (2017). Whole genome detection of signature of positive selection in African cattle reveals selection for thermotolerance. *Anim. Sci. J.* 88 (12), 1889–1901. doi: 10.1111/asj.12851

Tontis, A., Gutzwiller, A., and Zwahlen, R. (1992). Myocardial fibrosis and degeneration with heart failure (cardiomyopathy) in two goats. *Tierarztl. Prax.* 20 (4), 368–372.

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., et al. (2014). Design and characterization of a 52K SNP chip for goats. *PLoS One* 9 (1), e86227. doi: 10.1371/journal.pone.0086227

Tresset, A., and Vigne, J.-D. (2011). Last hunter-gatherers and first farmers of Europe. *C. R. Biol.* 334 (3), 182–189. doi: 10.1016/j.crvi.2010.12.010

Vahidi, S. M. F., Tarang, A. R., Anbaran, M. F., Boettcher, P., Joost, S., Colli, L., et al. (2014). Investigation of the genetic diversity of domestic Capra hircus breeds reared within an early goat domestication area in Iran. *Genet. Sel. Evol.* 46, 27. doi: 10.1186/1297-9686-46-27

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, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Vatsiou, A. I., Bazin, E., and Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol. Ecol.* 25 (1), 89–103. doi: 10.1111/mec.13360

Walugembe, M., Bertolini, F., Dematawewa, C. M. B., Reis, M. P., Elbeltagy, A. R., Schmidt, C. J., et al. (2018). Detection of selection signatures among Brazilian, Sri Lankan, and Egyptian chicken populations under different environmental conditions. *Front. Genet.* 9, 737. doi: 10.3389/fgene.2018.00737

Wang, X., Liu, J., Zhou, G., Guo, J., Yan, H., Niu, Y., et al. (2016). Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Sci. Rep.* 6, 38932. doi: 10.1038/srep38932

Weerasooriya, K. M. S. G., Fernando, P. S., Liyanagunawardena, N., Wijewardena, G., Wijemuni, M. I., and Samarakoon, S. A. T. C. (2017). Natural resistance of Sri Lankan village chicken to Salmonella gallinarum infection. *Br. Poult. Sci.* 58 (6), 644–648. doi: 10.1080/00071668.2017.1376034

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

Weldenegodguad, M., Popov, R., Pokharel, K., Ammosov, I., Ming, Y., Ivanova, Z., et al. (2018). Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. *Front. Genet.* 9, 728. doi: 10.3389/fgene.2018.00728

Wilde, C. G., Seilhamer, J. J., McGrogan, M., Ashton, N., Snable, J. L., Lane, J. C., et al. (1994). Bactericidal/permeability-increasing protein and lipopolysaccharide (LPS)-binding protein. LPS binding properties and effects on LPS-mediated cell activation. *J. Biol. Chem.* 269 (26), 17411–17416.

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21 (9), 1859–1875. doi: 10.1093/bioinformatics/bti310

Yang, J., Li, W. R., Lv, F. H., He, S. G., Tian, S. L., Peng, W. F., et al. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33 (10), 2576–2592. doi: 10.1093/molbev/msw129

Zeder, M. A., and Hesse, B. (2000). The initial domestication of goats (Capra hircus) in the Zagros Mountains 10,000 years ago. *Science* 287 (5461), 2254–2257. doi: 10.1126/science.287.5461.2254

Zeder, M. A. (2005). A view from the Zagros: new perspectives on livestock domestication in the Fertile Crescent, in *The First steps of Animal Domestication:*

New Archaeozoological Approcahes. Eds. J.-D. Vigne, J. Peters, and D. Helmer (Oxford, FL: Oxbow Books), 125–146.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 September 2019
doi: 10.3389/fgene.2019.00843

Optimal Management of Genetic Diversity in Subdivided Populations

Eugenio López-Cortegano[1,2], Ramón Pouso[1,2], Adriana Labrador[1,2], Andrés Pérez-Figueroa[1†], Jesús Fernández[3] and Armando Caballero[1,2]*

[1] Departamento de Bioquímica, Genética e Inmunología, Universidade de Vigo, Vigo, Spain, [2] Centro de Investigación Marina (CIM-UVIGO), Universidade de Vigo, Vigo, Spain, [3] Departamento de Mejora Genética, Instituto de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain

One of the main objectives of conservation programs is the maintenance of genetic diversity because this provides the adaptive potential of populations to face new environmental challenges. Genetic diversity is generally assessed by means of neutral molecular markers, and it is usually quantified by the expected heterozygosity under Hardy-Weinberg equilibrium and the number of alleles per locus or allelic diversity. These two measures of genetic diversity are complementary because whereas the former is directly related to genetic variance for quantitative traits and, therefore, to the short-term response to selection and adaptation, the latter is more sensitive to population bottlenecks and relates more strongly to the long-term capacity of adaptation of populations. In the context of structured populations undergoing conservation programs, it is important to decide the optimum management strategy to preserve as much genetic diversity as possible while avoiding inbreeding. Here we examine, through computer simulations, the consequences of choosing a conservation strategy based on maximizing either heterozygosity or allelic diversity of single-nucleotide polymorphism haplotypes in a subdivided population. Our results suggest that maximization of allelic diversity can be more efficient in maintaining the genetic diversity of subdivided populations than maximization of expected heterozygosity because the former maintains a larger number of alleles while making a better control of inbreeding. Thus, maximization of allelic diversity should be a recommended strategy in conservation programs for structured populations.

Keywords: conservation genetics, population management, allelic diversity, heterozygosity, genetic markers, SNP, haplotypes

OPEN ACCESS

Edited by:
Christos Palaiokostas, Swedish University of Agricultural Sciences, Sweden

Reviewed by:
Kevin Feldheim, Field Museum of Natural History, United States
Gregor Gorjanc, University of Edinburgh, United Kingdom

****Correspondence:***
Armando Caballero
armando@uvigo.es

†Present address:
Andrés Pérez-Figueroa, CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *20 April 2019*
Accepted: *13 August 2019*
Published: *13 September 2019*

Citation:
López-Cortegano E, Pouso R, Labrador A, Pérez-Figueroa A, Fernández J and Caballero A (2019) Optimal Management of Genetic Diversity in Subdivided Populations. Front. Genet. 10:843. doi: 10.3389/fgene.2019.00843

INTRODUCTION

Genetic diversity is the fuel for the adaptation of species to the environmental challenges and one of the main control variables to be assessed within the planetary boundaries framework (Steffen et al., 2015). Conservation of genetic diversity is also one of the main objectives for guaranteeing the long-term survival of species or breeds at risk of extinction (Frankham et al., 2010; Allendorf et al., 2013; Oldenbroek, 2017). Genetic diversity is generally assessed by means of neutral molecular markers in population genetics and conservation biology studies (Toro et al., 2009; Kirk and Freeland, 2011; Allendorf et al., 2013), and it is usually measured by the expected heterozygosity under Hardy-Weinberg equilibrium (Nei, 1973) and by the number of alleles per locus for multiallelic markers or allelic diversity. These two measures of genetic diversity are complementary because whereas the former is directly related to genetic variance for quantitative traits and, therefore, to the short-term

response to selection and adaptation for these traits (Falconer and Mackay, 1996), the latter is more sensitive to population bottlenecks (Luikart et al., 1998; Leberg, 2002), being thus useful to monitor them, and relates more strongly to the long-term response to natural and artificial selection (James, 1970; Hill and Rasbash, 1986; Wilson et al., 2009; Medugorac et al., 2011; Caballero and García-Dorado, 2013; Vilas et al., 2015). In the case of structured populations, subpopulation differentiation is traditionally measured through differences in gene frequency of alleles (Wright, 1952), but alternative ways to measure it based on allelic diversity have also been proposed (Petit et al., 1998; Jost, 2008; Caballero and Rodríguez-Ramilo, 2010; Jost et al., 2017).

The consensus criterion for the maintenance of genetic diversity in conservation and animal breeding programs is the maximization of expected heterozygosity, which is equivalent to the minimization of mean weighted coancestry (Toro and Pérez-Enciso, 1990; Ballou and Lacy, 1995; Meuwissen, 2007) and implies the maximization of effective population size (Caballero and Toro, 2000, Caballero and Toro, 2002). However, allelic diversity has also been proposed to establish conservation priorities (Bataillon et al., 1996; Petit et al., 1998; Fernández et al., 2004; Simianer, 2005; Caballero and Rodríguez-Ramilo, 2010; Medugorac et al., 2011; Jost et al., 2017; Ramljak et al., 2018; Zhang et al., 2018), and there is an increasing number of methods and computer tools developed to estimate and predict allelic richness (Belkhir et al., 2006; Szpiech et al., 2008; Bashalkhanov et al., 2009) and to retain the largest allelic diversity in conservation programs (Fernández et al., 2004; Weiser et al., 2012; López-Cortegano et al., 2019). Microsatellite analysis has also revealed that allelic richness is a better proxy for genome-wide single-nucleotide polymorphism (SNP) diversity than expected heterozygosity (Fischer et al., 2017). In addition, it has been argued that the number of allelic variants after a bottleneck might be the main factor responsible for the response to long-term selection and selection limits (p. 289) (Allendorf et al., 2013). In fact, Vilas et al. (2015) showed through experimental studies and simulation analyses that the long-term adaptive potential of a population is better indicated by allelic diversity than by expected heterozygosity.

Conservation programs can be aimed at maximizing either expected heterozygosity or allelic diversity. Fernández et al. (2004) compared these two alternative strategies for a single undivided population. In a set of simulations, populations were maintained over generations by choosing the parents' contributions to progeny that maximize the expected heterozygosity for multiallelic genetic markers. In another set, contributions were sought to maximize the number of marker alleles in the progeny. The results showed that each maximization method was, as would be expected, more efficient in maintaining each corresponding diversity measure. However, maximization of heterozygosity was able to maintain levels of allelic diversity almost as high as the method specifically devoted to that task. The explanation was that maximization of heterozygosity leads marker alleles toward intermediate frequencies because the maximal heterozygosity occurs when alleles are at equal frequencies. Thus, by spreading rare alleles to intermediate frequencies, their chances of loss by genetic drift are reduced. A method specifically focused on keeping allelic diversity was effective in doing so but some rare alleles were maintained at low frequencies, being more likely to be eventually lost.

The results from Fernández et al. (2004) were carried out in the context of a single undivided population. Most populations, however, in nature and in conservation programs (zoos, germplasm collections, botanic gardens, etc.) are subdivided. As suggested by preliminary analyses, the outcomes of maximizing expected heterozygosity or allelic diversity could be very different in subdivided populations (López-Cortegano et al., 2019). Thus, a question arises as to which of these methods is more efficient in maintaining genetic diversity while controlling inbreeding in structured populations. Here, we address this issue by performing simulations of a subdivided population and a conservation program where maximization of heterozygosity and allelic diversity are carried out for two sets of genetic markers, one representing a small number of known loci where diversity should be preserved and another aimed to perform whole-genome management. Because of the increasing availability of genotyping and sequencing projects, we focus on haplotypic combinations of SNPs as the marker of choice for future conservation strategies.

MATERIALS AND METHODS

Simulations were carried out in two steps. In the first, individual-based forward simulations were run to generate a subdivided population (**Figure 1**). An ancestral large population of 4,000 individuals was first run for 5,000 generations to build sufficient neutral genetic variation under a mutation-drift equilibrium. From this large base population, five subpopulations were founded, one of size $N = 2{,}000$ and four of size $N = 100$ individuals, to obtain different degrees of variation within subpopulations, which were maintained independently for 25 generations of

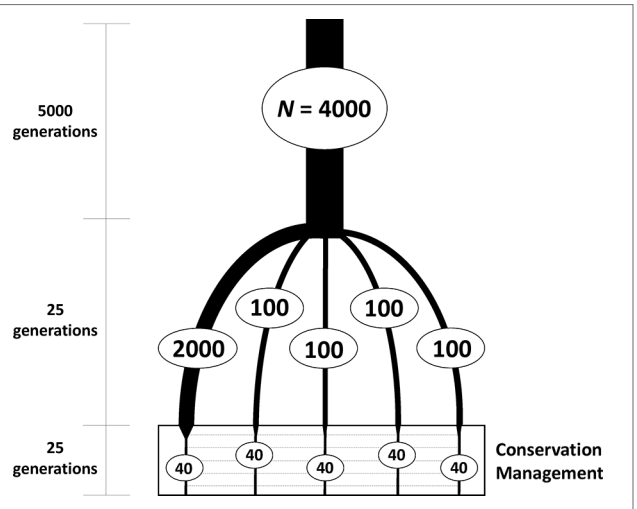


FIGURE 1 | Scheme of the evolutionary history of the simulated subdivided population to be used as a base for a conservation program. One ancestral population of large size (N = 4,000 diploid individuals) was first maintained for a long period of time to reach mutation-drift equilibrium. From this ancestral population, five subpopulations (with constant population sizes as shown in the figure) were founded and maintained independently for 25 generations. These subpopulations were thereafter maintained with 40 individuals each and subjected to a conservation program aimed at maximizing either expected heterozygosity or allelic diversity of neutral markers, with some migration allowed between subpopulations.

random mating. The software SLiM 3 (Haller and Messer, 2018) was used as a forward genomic simulator in this first step. All simulations involved a sequence of 10 Kb with mutation rate of 5×10^{-5} per nucleotide and generation and a recombination rate of 10^{-6} between consecutive nucleotides in the formation of gametes. These mutation and recombination rate values were chosen to obtain a sufficiently high number and density of polymorphic loci within the simulated sequence. Additional simulations were also performed assuming a recombination rate one order of magnitude higher. Random mating of parents under the assumption of neutrality was implemented.

The second step of the study was the conservation management of the structured population created from the previous simulation. From each of the five subpopulations, a sample of $N = 40$ individuals (20 of each sex) was obtained and maintained with that size under a common conservation scheme based on maximization of either expected heterozygosity or allelic diversity for 25 generations with controlled migration between subpopulations. Marker loci to be used for analysis and management were assumed to be haplotypes of groups of five consecutive SNPs, such that the different haplotypic combinations of SNPs per locus were considered as different alleles, providing a maximum of 32 per locus. The total number of available loci was about 2,000, but the number of segregating loci available for analysis at the start of the conservation management process was approximately 1,200.

For conservation management, we used the software Metapop2 (López-Cortegano et al., 2019), which provides in each generation the optimal mating crosses and contributions from parents to the next generation as well as the number and specific migrants across subpopulations to maximize either heterozygosity or allelic diversity with a control in the number of migrations. With this program, total heterozygosity (H_T) is partitioned, following Nei (1973), in the average expected heterozygosity within subpopulations assuming Hardy-Weinberg proportions (H_S) and the average Nei's minimum genetic distance between subpopulations, averaged over all possible pairs of subpopulations (D_G). In an analogous way, allelic diversity is partitioned, following Caballero and Rodríguez-Ramilo (2010), in a within- and between-subpopulation component of variation. The within-subpopulation component is the average number of alleles segregating in the subpopulations minus one (A_S). The between-subpopulation component (D_A) is calculated as the number of alleles present in a subpopulation and absent in other when subpopulations are compared in pairs and averaged over all possible pairs of subpopulations. Total allelic diversity is then defined as $A_T = A_S + D_A$ and represents the total number of alleles present in a given pair of subpopulations, averaged for all possible pairs.

The Metapop2 software performs an optimization of contributions of parents to progeny and migrations between subpopulations with the dynamic method of Fernández et al. (2008) to maximize diversity. Maximization of total expected heterozygosity (maxH_T) or total allelic diversity (maxA_T) is obtained by maximizing the functions $H_T = D_G + \lambda H_S$ and $A_T = D_A + \lambda A_S$, respectively, where λ is the desired weight given to the within-subpopulation component. In addition, the program also maximizes the total allelic number in the whole population (maxK) by managing contributions from parents to progeny and migrations so that the global probability of alleles' losses in the progeny is minimized (Vales-Alonso et al., 2003; Fernández et al., 2004). Note that maxK pursues a maximization of the total number of alleles in the population without regard to the distribution of these across subpopulations. Because a maximum number of alleles in the whole population would be obtained with a maximum differentiation between subpopulations, maxK is expected to lead to such a situation. Maximization of A_T, in contrast, implies a control on the distribution of the alleles maintained across subpopulations particularly if different weights are given to the within- and between-subpopulation components of diversity. Thus, alleles can be conserved uniformly distributed, leading to a reduction in the differentiation between subpopulations, or variably distributed across subpopulations, leading to an increase in the differentiation (López-Cortegano et al., 2019). At one extreme, each allele of a locus could be maintained in a different subpopulation. At the other, all different alleles for a locus could be maintained simultaneously in all subpopulations.

Management was run assuming two different objectives: (1) Conservation of diversity for a particular set of loci for which one locus every 100 in the genome was used for management and genetic variation was measured directly on that set of loci. This refers to a scenario in which a few known loci or genomic regions of particular interest have to be managed, for example, for loci that are known to have an effect on a particular trait of interest, such as those affecting a productive trait, the immune system, and so on. (2) Conservation of diversity in the whole genome for which one locus every 10 was used for management and the results were analyzed for all genomic loci. This is a situation where a number of markers are used for conserving overall genetic diversity. For this latter case, we used a modification of the software Metapop2 (López-Cortegano et al., 2019). With the assumed simulated sequence length and recombination rates, the density of markers would be in the range between 1,200 and 12,000 per Morgan, thus implying a high marker density in prevision of the increasing availability of dense SNP chips for more and more species. In the management period, it was assumed that there was no recombination within loci (*i.e.,* between SNPs of a particular haplotype) and recombination was free between them, which are reasonable assumptions given the short number of SNPs per locus and the scarcity of loci used along the sequence.

The optimization was carried out for 25 generations generally assuming a value of $\lambda = 1$, thus giving the same weight to within- and between-subpopulation components. However, other values of λ were also considered, including those for which all weight is given to between-subpopulation diversity ($\lambda = 0$), all weight is given to within-subpopulation diversity ($\lambda = 1000$), and $\lambda = 0.5$, a value suggested to maximize the total genetic variance of a hypothetical quantitative trait (Bennewitz and Meuwissen, 2006). A maximum possible number of migrants of one per subpopulation and generation were assumed, a typical rule of thumb suggested to maintain a considerable differentiation between subpopulations but avoiding an excessive increase in inbreeding (Mills and Allendorf, 1996). In all cases, 10 replicates of the base population were simulated and, for each of them, 10 different sampling events and management processes were run. In every generation, the

average values over replicates of expected heterozygosity measures (H_T, H_S and D_G), allelic diversity measures (A_T, A_S, D_A and K), and the observed marker homozygosity of all individuals in the subpopulations (to which we will refer to as molecular inbreeding, F, and which includes homozygotes identical by descent and identical in state) were obtained from the 12 managed markers for the scenario aimed at conserving diversity for a specific set of loci and from the whole sequence in the scenario aimed at conserving diversity for the whole genome.

RESULTS

Three optimization methods were compared (maxH_T, maxA_T, and maxK), aimed at maximizing global heterozygosity H_T, global allelic diversity A_T, and the total number of alleles K, respectively. The evolution of these parameters and their within- and between-subpopulation components are shown in **Figures 2** and **3** when the same weight is given to within- and between-subpopulation diversity ($\lambda = 1$). As expected, no management (RND; black dotted lines) led to a generalized loss of genetic diversity and to an increase in molecular inbreeding whereas any of the specific management methods increased diversity or restrained its loss through generations. The relative performance of the different optimization methods was very similar for scenarios aiming at the conservation of diversity for either a particular set of loci (**Figure 2**) or the whole genome (**Figure 3**). Thus, we describe them simultaneously.

As expected, each maximization method maintained higher levels of the corresponding measure of diversity. Thus, maxH_T (blue lines) was the best method, preserving H_T in the global population by means of an initial increase in the diversity between subpopulations (D_G) while keeping or slightly decreasing

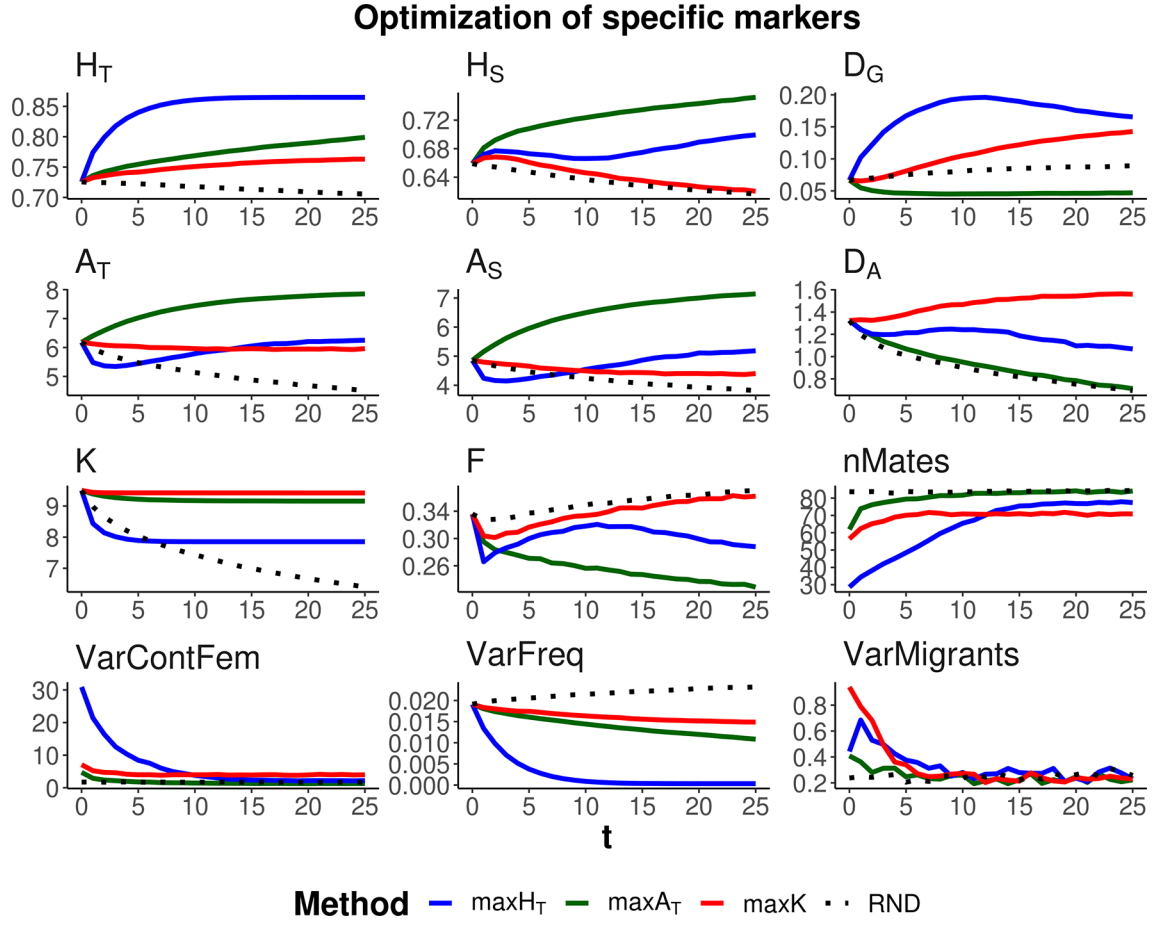


FIGURE 2 | Changes in different diversity parameters over generations (t) in a subdivided population subjected to three optimization methods, maxH_T (blue line), maxA_T (green line), and maxK (red line), and an unmanaged control (RND, dotted black line). Optimization was made for a particular set of loci (12 multiallelic single-nucleotide polymorphism haplotype markers). In the case of maxH_T and maxA_T, a between-population weighting factor of $\lambda = 1$ (*i.e.,* equal weight for within- and between-subpopulation components of diversity) was assumed. One migrant per subpopulation and generation was considered in the optimizations. Statistics measured in the managed markers: total heterozygosity (H_T) and its within- and between-subpopulation components (H_S and D_G); total allelic diversity (A_T) and its within- and between-subpopulation components (A_S and D_A); total number of alleles in the population (K); average coefficient of molecular inbreeding of individuals (F); number of pairing mates involved in the different procedures (nMates); variance of the contribution from female parents to progeny (varContFem); variance of allelic frequencies with loci (VarFreq); and variance of the number of migrants per subpopulation (VarMigrants). Standard errors for means are lower than 0.01 (allelic measures), 0.001 (heterozygosity measures), and 0.002 (F).

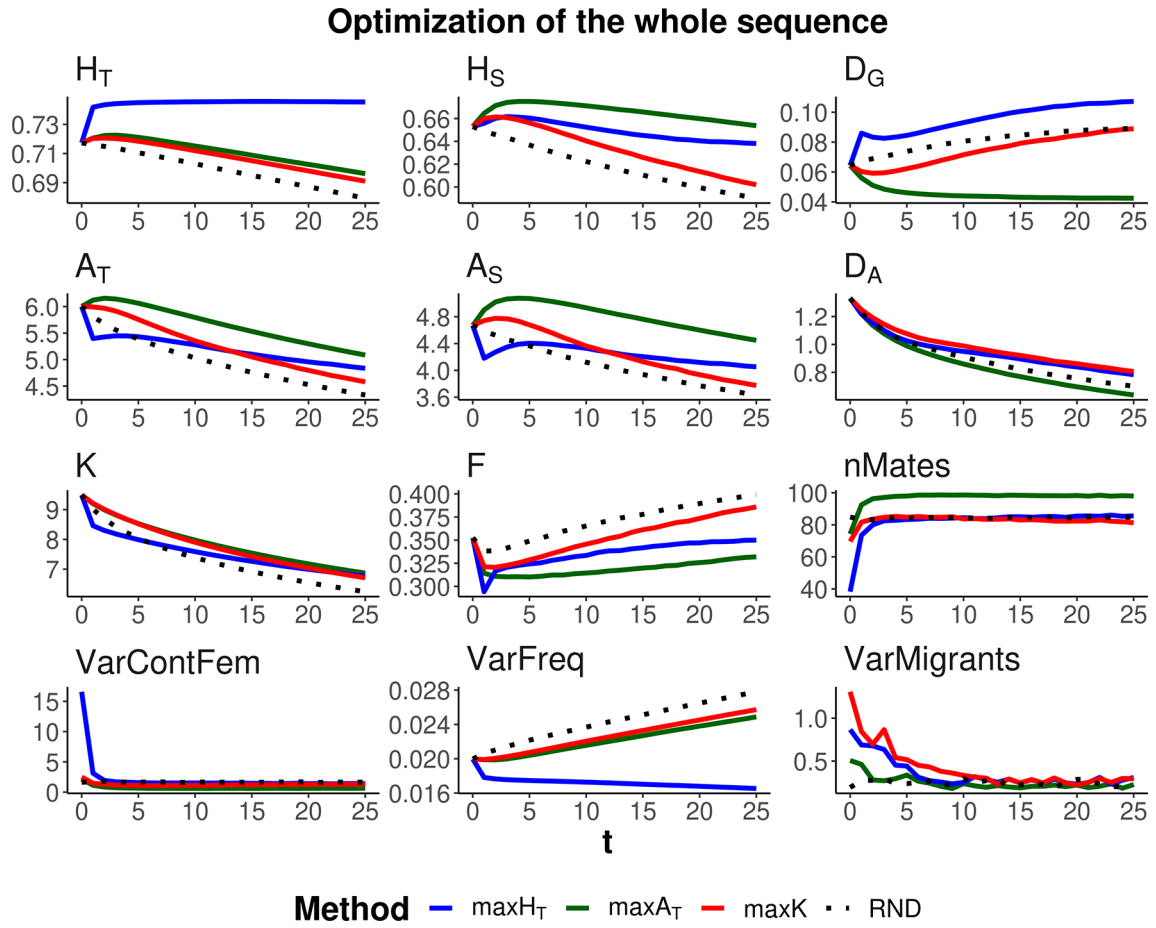


FIGURE 3 | Changes in different diversity parameters over generations (*t*) in a subdivided population subjected to three optimization methods, maxH$_T$ (blue line), maxA$_T$ (green line), and maxK (red line), and an unmanaged control (RND, dotted black line). Optimization was made for 120 multiallelic (single-nucleotide polymorphism haplotype) markers, but statistics were calculated for the whole simulated genome. Simulation characteristics and statistics as in **Figure 2**. Standard errors for means are lower than 0.01 (allelic measures), 0.001 (heterozygosity measures), and 0.002 (*F*).

that within subpopulations (H_S). Method maxA$_T$ (green lines) produced the largest A_T by increasing or keeping a high diversity within subpopulations (A_S) and decreasing that between subpopulations (D_A). Finally, maxK (red lines) maintained the largest number of alleles segregating in the whole population (K), although maxA$_T$ maintained only a little less or about the same number of alleles.

The molecular inbreeding coefficient (F) was better restrained by maxA$_T$, whereas maxK produced the highest molecular inbreeding levels, close to those yielded by RND. Method maxA$_T$ was also the optimizing method making a wider use of the individuals available for mating (nMates) and produced the lowest variance of contributions from females to progeny (VarContFem), thus approaching the equalization of contributions from parents to progeny. Method maxH$_T$ produced the highest variance of contributions from females to the progeny in the initial generations.

As already observed in previous studies, maxH$_T$ tends to equalize allele frequencies within loci to reach the maximum possible heterozygosity. This can be seen as a reduction in the variance of allelic frequencies within loci (VarFreq). Method maxK was producing the largest variation in allelic frequencies.

All management methods involved an average number of one migrant per generation and subpopulation. However, there were differences in the variance of the number of migrants per subpopulation depending on the generations and methods (see graphs VarMigrants in **Figures 2** and **3**). The highest variation occurred in the initial generations when differences in diversity between subpopulations were larger. Most migrations in these initial generations occurred from the first subpopulation (that with the largest ancestral size; **Figure 1**) to the others (not shown). Method maxA$_T$ was the optimizing procedure with the lowest variation in subpopulation migrations.

Figure 4 shows the results corresponding to the scenario of conservation of diversity for a particular set of loci (the same as in **Figure 2**) for a range of values of the weight (λ) given to the within-subpopulation component. Method maxK and no management (RND) obviously were unaffected by the different weighting. Method maxH$_T$ maintained the highest H_T for all λ

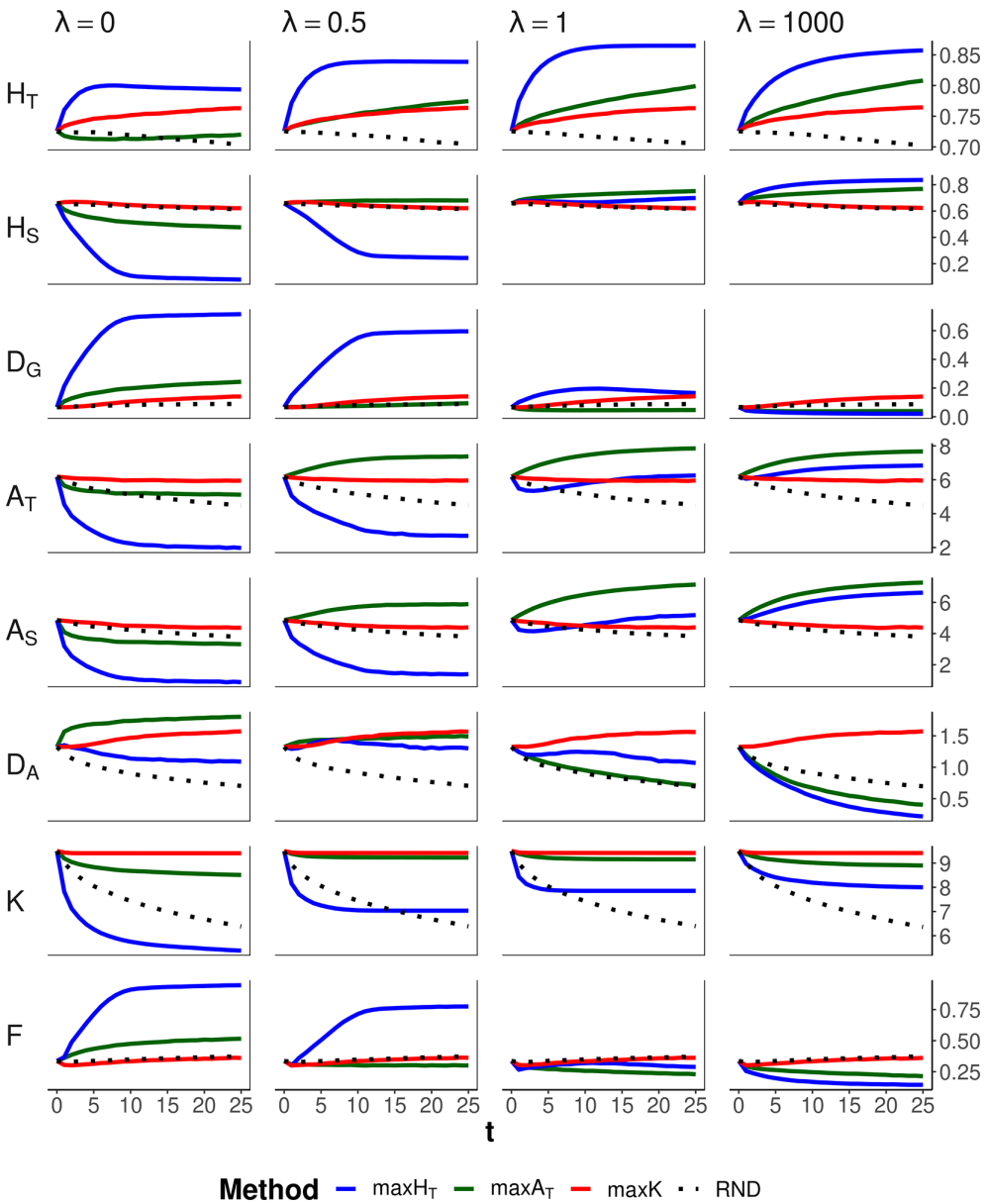


FIGURE 4 | Changes in different diversity parameters over generations (t) in a subdivided population subjected to three optimization methods, maxH_T (blue line), maxA_T (green line), and maxK (red line), and an unmanaged control (RND, dotted black line). Optimization was made for a particular set of loci (12 multiallelic single-nucleotide polymorphism haplotype markers). In the case of maxH_T and maxA_T, different between-population weighting factors (λ) were assumed. One migrant per subpopulation and generation was considered in the optimizations. Statistics refer to the managed markers: total heterozygosity (H_T) and its within- and between-subpopulation components (H_S and D_G); total allelic diversity (A_T) and its within- and between-subpopulation components (A_S and D_A); total number of alleles in the population (K); and average coefficient of molecular inbreeding of individuals (F). Standard errors for means are lower than 0.01 (allelic measures), 0.001 (heterozygosity measures), and 0.002 (F).

values. This was attained by increasing D_G when all weight is given to the between-subpopulation component (λ = 0) at the expense of decreasing the within-subpopulation component H_S, or by increasing H_S when all weight is given to the within-subpopulation component (λ = 1,000) at the expense of decreasing the between-subpopulation component D_G. Method maxA_T preserved better A_T when some substantial weight was given to the within-subpopulation component (*i.e.*, $\lambda \geq 0.5$). If all weight is given to the within-subpopulation component (λ = 1,000), maxH_T would produce the lowest molecular inbreeding (F), as expected, but the number of alleles maintained would be lower than those obtained by the allelic optimization methods. For intermediate values of λ (0.5 or 1), maxA_T seems to be the most robust method, producing the lowest inbreeding and a number of alleles almost as large as that maintained by maxK, although giving lower H_T than that of maxH_T.

Additional simulations regarding alternative parameter settings with respect to those considered above are given as Supplementary figures. First, the results shown in **Figures 2–4** involved an average of one migrant per generation and subpopulation in the management period. **Supplementary Figure S1** presents results analogous to those of **Figure 2** but including a lower (0.4) and a higher (2) average number of migrations per subpopulation and generation, showing that the main results basically hold. Finally, **Figures 2–4** refer to simulations with a recombination rate between nucleotides of $c = 10^{-6}$. **Supplementary Figures S2 and S3** show results analogous to those of **Figures 2** and **3** but considering a recombination rate one order of magnitude larger ($c = 10^{-5}$). The results are, in general terms, also similar to those obtained before.

DISCUSSION

Preservation of genetic diversity is one of the main objectives of conservation programs (Frankham et al., 2010; Allendorf et al., 2013; Oldenbroek, 2017). Because many threatened species have fragmented habitats and many populations maintained in captivity are structured, conservation methods should consider population subdivision and focus on a global management, including possible migrations among subpopulations, rather than being restricted to local efforts (Frankham et al., 2010, Chap. 17). In addition, in the absence of genealogical data, molecular markers are used to analyze population diversity and make conservation designs regarding genetic objectives (Benestan et al., 2016; Fuentes-Pardo and Ruzzante, 2017). Here we have addressed the question of which marker diversity parameters should be better considered for making conservation management decisions in a subdivided population. For multiallelic markers (such as microsatellite loci) or biallelic ones (such as SNPs) that can be analyzed as multiallelic ones if considering multi-SNP haplotypes (*e.g.,* Zhao et al., 2019), decisions can be taken on expected heterozygosity or allelic diversity measures. We have investigated the outcome of a subdivided population maintained with different optimization procedures aimed at maximizing heterozygosity or allelic diversity. Each method was successful in maintaining the diversity measure aimed at, but they showed remarkable differences on how much of the rest of diversity parameters are conserved, the distribution patterns of diversity within and between populations, and the level of molecular inbreeding (homozygosity). The results confirm some preliminary runs carried out by López-Cortegano et al. (2019) to illustrate the use of the software Metapop2 with multiallelic markers. Thus, allelic diversity methods, in particular maxA_T, can be recommended as the method of choice because it maintains a high allelic richness in the population (uniformly distributed across subpopulations) and controls inbreeding rather efficiently.

We considered two scenarios regarding the number of markers to be managed. One in which a specific set of loci is the target for conservation, as it could apply, for example, to specific loci of interest, such as those related to the immune system. In this case, because management is carried out on the specific loci of interest, the management methods are very effective in increasing genetic diversity (**Figure 2**). Another scenario has the objective of preserving the whole genomic diversity by using a restricted number of markers. In this case, the management methods are obviously less effective (**Figure 3**), and the degree of success will depend on the number of markers considered and the genetic structure of the species. We used a relatively high density of markers and, in this situation, the methods were rather effective in conserving genetic diversity for the whole sequence. However, it is expected that the availability of only a low number of markers will be less effective in achieving proper management of the whole genome.

It has been suggested that the number of alleles relates more strongly to the long-term capacity of populations to adapt to changing environments (James, 1970; Hill and Rasbash, 1986; Wilson et al., 2009; Medugorac et al., 2011). Caballero and García-Dorado (2013) showed, through computer simulations, that the long-term adaptive potential of a subdivided population subject to natural selection relates more strongly to allelic diversity. Vilas et al. (2015) performed an experiment with *Drosophila melanogaster* in which synthetic populations were built from a group of subpopulations by maximizing either the heterozygosity or the total number of alleles for nine microsatellite loci. Artificial selection for sternopleural bristle number during eight generations showed that the response to selection was larger (for both upward and downward number of bristles) for synthetic populations obtained by maximizing the number of marker alleles than for those obtained by maximizing marker heterozygosity. In addition, it has been observed in *Arabidopsis halleri* that genome-wide SNP diversity does not show a significant correlation with microsatellite heterozygosity based on 20 markers but is significantly correlated with microsatellite allelic richness (Fischer et al., 2017). These results thus suggest that maximization of allelic diversity can be a more desirable conservation strategy than maximization of expected heterozygosity of multiallelic markers regarding the maintenance of the adaptive potential of populations. On the other hand, inbreeding must also be avoided because of the negative effects associated to inbreeding depression (Charlesworth and Willis, 2009). Method maxA_T seems to accomplish both objectives.

Maximizing global heterozygosity is achieved by leading genes to intermediate allele frequencies (Fernández et al., 2004). In fact, maximizing heterozygosity is equivalent to maximizing the effective number of alleles, that is, the number of alleles per locus if all had the same frequency (Crow and Kimura, 1970). We checked this by performing simulations where a global optimization is made on the total effective number of alleles in the population, finding results identical to those for maxH_T with $\lambda = 1$. In a single undivided population, this tendency for equalizing allelic frequencies within each locus has the advantage of leading rare alleles to intermediate frequencies and thus also avoiding their loss. Thus, in undivided populations, maxH_T can be the most appropriate method to be carried out for conserving both a high heterozygosity and a high number of alleles (Fernández et al., 2004). In subdivided populations, maxH_T also

implies a reduction in the variance of allelic frequencies within loci (VarFreq in **Figures 2** and **3**), but maximization of global heterozygosity is made at the cost of an increase of homozygosity (and thus inbreeding) in each subpopulation, at least in the short term, and a substantial loss of alleles (*F* and *K*, respectively, in **Figures 2** and **3**). Only if all weight in the optimization is given to within-subpopulation variation, maxH_T would make the best control of molecular inbreeding (λ = 1,000 in **Figure 4**). However, the overall number of alleles maintained would still be lower than that maintained by the allelic diversity optimization methods (*K* in **Figure 4**).

Regarding allelic diversity procedures, we have compared the allelic diversity partition suggested by Caballero and Rodríguez-Ramilo (2010) (maxA_T) with a method aimed at maintaining the overall number of alleles in the population (maxK). Although the former can be used to control the distribution of allelic variants within and between subpopulations, the second is applied without such a control. A notable different outcome is observed with each method. Method maxK maximizes, as expected, the total number of alleles in the whole population, but alleles are distributed variably across subpopulations, as indicated by a high value of D_A. In contrast, maxA_T maintains almost as many alleles as maxK in the whole population but keeps them more homogeneously distributed over subpopulations, as indicated by a low value of D_A. In conservation programs of structured populations, the objective may be to maintain reservoirs of variation such that there is little overlap between different subpopulations, for example, when there are local adaptations and a risk of outbreeding depression, in which case, a method such as maxK could be more appropriate. However, the loss of a subpopulation implies, in this case, the irreversible loss of allelic variation. If, on the contrary, allelic diversity is maintained uniformly in all subpopulations, as achieved by maxA_T (and, to some extent, by maxH_T), the loss of a subpopulation does not imply a loss of allelic diversity because each subpopulation would provide a backup for the others. In a recent article, Ramljak et al. (2018) have proposed to use the statistic A_T to prioritize different European cattle breeds for conservation.

Ollivier and Foulley (2013) have argued that the partition of allelic diversity proposed by Caballero and Rodríguez-Ramilo (2010) does not meet two properties. First, that the partition of within- and between-subpopulation components is not orthogonal because both components are not independent. Second, that it does not meet concavity, which means that diversity cannot decrease when a subpopulation is added or increase when a subpopulation is dropped. The lack of these supposedly desirable properties also affects Nei's heterozygosity partition because both partitions follow the same approach. The lack of orthogonality of Nei's partition has also been discussed by Jost (2008) (but see also Whitlock, 2011; Wang, 2012). Ollivier and Foulley (2013) recommended a definition of allelic diversity that relies mostly on the presence of private alleles, that is, a subpopulation only contributes to the total allelic diversity if it carries unique alleles in the population. Thus, if the subpopulations have no private alleles, their contribution to global allelic diversity is null and, in that scenario, the distribution of the allelic variants across subpopulations is irrelevant. In that sense, method maxK, whose objective is to maximize the total number of allelic variants in the whole population, would be consistent with that view of managing allelic diversity. Our results, in fact, show that maxK maximizes the total number of alleles, but maxA_T produces almost the same outcome in terms of total allelic number with the desirable addition of a better control of inbreeding.

In summary, our results suggest that maxA_T, the maximization of the total allelic diversity (A_T) following Caballero and Rodríguez-Ramilo (2010), which represents the total number of alleles present in a given pair of subpopulations averaged for all possible pairs, could be recommended as a standard management method for conservation programs of structured populations on the basis that it is efficient in preserving allelic diversity, within-subpopulation variation, and restraining inbreeding, thus guaranteeing the capacity of adaptation to short- and long-term environmental challenges.

DATA AVAILABILITY

The pipeline required to simulate all populations assayed as well as to run the corresponding population analyses and management simulations is available in https://gitlab.com/elcortegano/hadopt.

AUTHOR CONTRIBUTIONS

AC, EL-C, and JF contributed to the conception and design of the study; EL-C and AP-F developed the required software and pipelines; EL-C, RP, and AL performed computer simulations; AC and EL-C wrote the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This work was supported by Agencia Estatal de Investigación (AEI) (CGL2016-75904-C2), Xunta de Galicia (ED431C 2016-037), and Fondos Feder: "Unha maneira de facer Europa." UVigo Marine Research Centre (CIM-UVIGO) is funded by the "Excellence in Research (INUGA)" Program from the Regional Council of Culture, Education and Universities, with co-funding from the European Union through the ERDF Operational Program Galicia 2014-2020.

ACKNOWLEDGMENTS

We thank three reviewers who helped improve the manuscript. The analyses reported here were performed on the Finis Terrae machine provided by CESGA (Galicia Supercomputing Centre).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00843/full#supplementary-material

REFERENCES

Allendorf, F. W., Luikart, G. H., and Aitken, S. N. (2013). *Conservation and the genetics of populations*. Chichester, West Sussex, UK: John Wiley and Sons.

Ballou, J. D., and Lacy, R. C. (1995). "Identifying genetically important individuals for management of genetic diversity in pedigreed populations," in *Population Management for Survival and Recovery*. Eds. J. D. Ballou, M. Gilpin, and T. J. Foose (New York: Columbia University Press), 76–111.

Bashalkhanov, S., Pandey, M., and Rajora, O. P. (2009). A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genet.* 10, 84. doi: 10.1186/1471-2156-10-84

Bataillon, T. M., David, J. L., and Schoen, D. J. (1996). Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics* 144, 409–417.

Belkhir, K., Dawson, K. J., and Bonhomme, F. (2006). A comparison of rarefaction and Bayesian methods for predicting the allelic richness of future samples on the basis of currently available samples. *J. Hered.* 97, 483–492. doi: 10.1093/jhered/esl030

Benestan, L. M., Ferchaud, A. L., Hohenlohe, P. A., Garner, B. A., Naylor, G. J. P., Baums, I. B., et al. (2016). Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Mol. Ecol.* 25, 2967–2977. doi: 10.1111/mec.13647

Bennewitz, J., and Meuwissen, T. H. E. (2006). Breed conservation priorities derived from contributions to the total future genetic variance. Proc. 8th World Congress on Genetics Applied to Livestock Production, CD-Rom Communication No. 9, 33–06.

Caballero, A., and García-Dorado, A. (2013). Allelic diversity and its implications for the rate of adaptation. *Genetics* 195, 1373–1384. doi: 10.1534/genetics.113.158410

Caballero, A., and Rodríguez-Ramilo, S. T. (2010). A new method for the partition of allelic diversity within and between subpopulations. *Conserv. Genet.* 11, 2219–2229. doi: 10.1007/s10592-010-0107-7

Caballero, A., and Toro, M. A. (2000). Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genet. Res.* 75, 331–343. doi: 10.1017/S0016672399004449

Caballero, A., and Toro, M. A. (2002). Analysis of genetic diversity for the management of conserved subdivided populations. *Conserv. Genet.* 3, 289–299. doi: 10.1023/A:1019956205473

Charlesworth, D., and Willis, J. H. (2009). The genetics of inbreeding depression. *Nat. Rev. Genet.* 10, 783–796. doi: 10.1038/nrg2664

Crow, J. F., and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.

Falconer, D. S., and Mackay, T. C. (1996). *Introduction to Quantitative Genetics*. 4th edn. Harlow, Essex, UK: Longmans Green.

Fernández, J., Toro, M. A., and Caballero, A. (2004). Managing individuals' contributions to maximize the allelic diversity maintained in small, conserved populations. *Conserv. Biol.* 18 (5), 1358–1367. doi: 10.1111/j.1523-1739.2004.00341.x

Fernández, J., Toro, M. A., and Caballero, A. (2008). Management of subdivided populations in conservation programs: Development of a novel dynamic system. *Genetics* 179, 683–692. doi: 10.1534/genetics.107.083816

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., et al. (2017). Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genom.* 18, 69. doi: 10.1186/s12864-016-3459-7

Frankham, R., Ballou, J. D., and Briscoe, D. A. (2010). *Introduction to Conservation Genetics*. Cambridge, UK: Cambridge university press. doi: 10.1017/CBO9780511809002

Fuentes-Pardo, A. P., and Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol. Ecol.* 26, 5369–5406. doi: 10.1111/mec.14264

Haller, B. C., and Messer, P. W. (2018). SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.* 36 (3), 632–637. doi: 10.1101/418657

Hill, W. G., and Rasbash, J. (1986). Models of long term artificial selection in finite populations. *Genet. Res.* 48, 41–50. doi: 10.1017/S0016672300024642

James, J. W. (1970). The founder effect and response to artificial selection. *Genet. Res.* 16, 241–250. doi: 10.1017/S0016672300002500

Jost, L. (2008). G_{ST} and its relatives do not measure differentiation. *Mol. Ecol.* 17, 4015–4026. doi: 10.1111/j.1365-294X.2008.03887.x

Jost, L., Archer, F., Flanagan, S., Gaggiotti, O., Hoban, S., and Latch, E. (2017). Differentiation measures for conservation genetics. *Evol. Appl.* 11, 1139–1148. doi: 10.1111/eva.12590

Kirk, H., and Freeland, J. R. (2011). Applications and implications of neutral versus non-neutral markers in molecular ecology. *Int. J. Mol. Sci.* 12, 3966–3988. doi: 10.3390/ijms12063966

Leberg, P. L. (2002). Estimating allelic richness: effects of sample size and bottlenecks. *Mol. Ecol.* 11, 2445–2449. doi: 10.1046/j.1365-294X.2002.01612.x

López-Cortegano, E., Pérez-Figueroa, A., and Caballero, A. (2019). Metapop2: re-implementation of software for the analysis and management of subdivided populations using gene and allelic diversity. *Mol. Ecol. Resour.* 19, 1095–1100. doi: 10.1111/1755-0998.13015

Luikart, G., Allendorf, F., Cornuet, J. M., and Sherwin, W. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J. Hered.* 89, 238–247. doi: 10.1093/jhered/89.3.238

Medugorac, I., Veit-Kensch, C. E., Ramljak, J., Brka, M., Markovic, B., Stojanovic, S., et al. (2011). Conservation priorities of genetic in domesticated metapopulations: a study in taurine cattle breeds. *Ecol. Evol.* 1, 408–420. doi: 10.1002/ece3.39

Meuwissen, T. H. E. (2007). "Operation of conservation schemes," in *Utilisation and Conservation of Farm Animal Genetic Resources*. Ed. K. Oldenbroek (Wageningen, The Netherlands: Wageningen Academic Publishers), 167–193.

Mills, L. S., and Allendorf, F. W. (1996). The one-migrant-per-generation rule in conservation and management. *Conserv. Biol.* 10, 1509–1518. doi: 10.1046/j.1523-1739.1996.10061509.x

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Oldenbroek, J. K., editors. (2017). *Genomic management of animal genetic diversity*. The Netherlands: Wageningen Academic Publishers. doi: 10.3920/978-90-8686-850-6

Ollivier, L., and Foulley, J.-L. (2013). A note on the partition of allelic diversity. *Conserv. Genet.* 14, 1285–1290. doi: 10.1007/s10592-013-0508-5

Petit, R. J., El Mousadik, A., and Pons, O. (1998). Identifying populations for conservation on the basis of genetic markers. *Conserv. Biol.* 12, 844–855. doi: 10.1046/j.1523-1739.1998.96489.x

Ramljak, J., Bunevski, G., Bytyqi, H., Marković, B., Brka, M., Ivanković, A., et al. (2018). Conservation of a domestic metapopulation structured into related and partly admixed strains. *Mol. Ecol.* 27, 1633–1650. doi: 10.1111/mec.14555

Simianer, H. (2005). Using expected allele number as objective function to design between and within breed conservation of farm animal biodiversity. *J. Anim. Breed Genet.* 122, 177–187. doi: 10.1111/j.1439-0388.2005.00523.x

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., et al. (2015). Sustainability. Planetary boundaries: guiding human development on a changing planet. *Science* 347 (6223), 1259855. doi: 10.1126/science.1259855

Szpiech, Z. A., Jakobsson, M., and Rosenberg, N. A. (2008). ADZE: a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* 24, 2498–2504. doi: 10.1093/bioinformatics/btn478

Toro, M. A., Fernández, J., and Caballero, A. (2009). Molecular characterization of breeds and its use in conservation. *Livest. Sci.* 120, 174–195. doi: 10.1016/j.livsci.2008.07.003

Toro, M. A., and Pérez-Enciso, M. (1990). Optimization of selection response. *Genet. Sel. Evol.* 22, 93–107. doi: 10.1186/1297-9686-22-1-93

Vales-Alonso, J., Fernández, J., González-Castaño, F. J., and Caballero, A. (2003). A parallel optimization approach for controlling allele diversity in conservation schemes. *Math. Biosci.* 183, 161–173. doi: 10.1016/S0025-5564(03)00037-3

Vilas, A., Pérez-Figueroa, A., Quesada, H., and Caballero, A. (2015). Allelic diversity for neutral markers retains a higher adaptive potential for quantitative traits than expected heterozygosity. *Mol. Ecol.* 24 (7), 4419–4432. doi: 10.1111/mec.13334

Wang, J. (2012). On the measurements of genetic differentiation among populations. *Genet. Res.* 94, 275–289. doi: 10.1017/S0016672312000481

Weiser, E. L., Grueber, C. E., and Jamieson, I. G. (2012). AlleleRetain: a program to assess management options for conserving allelic diversity

in small, isolated populations. *Mol. Ecol. Resour.* 12, 1083–1091. doi: 10.1111/j.1755-0998.2012.03176.x

Whitlock, M. C. (2011). G'_{ST} and D do not replace F_{ST}. *Mol. Ecol.* 20, 1083–1091. doi: 10.1111/j.1365-294X.2010.04996.x

Wilson, A., Arcese, P., Keller, L. F., Pruett, C. L., Winker, K., Patten, M. A., et al. (2009). The contribution to island populations to *in situ* genetic conservation. *Conserv. Genet.* 10, 419–430. doi: 10.1007/s10592-008-9612-3

Wright, S. (1952). The theoretical variance within and among subdivisions of a population that is in a steady state. *Genetics* 37, 312–321.

Zhang, M., Peng, W. F., Hu, X. J., Zhao, Y. X., and Yang, J. (2018). Global genomic diversity and conservation priorities for domestic animals are associated with the economies of their regions of origin. *Sci. Rep.* 8 (1), 11677. doi: 10.1038/s41598-018-30061-0

Zhao, Q.-B., Sun, H., Zhang, Z., Xu, Z., Olasege, B. S., Ma, P.-P., et al. (2019). Exploring the structure of haplotype blocks and genetic diversity in Chinese indigenous pig populations for conservation purpose. *Evol. Bioinform. Online* 15, 1176934318825082 doi: 10.1177/1176934318825082.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 October 2019
doi: 10.3389/fgene.2019.00901

Multiple Selection Signatures in Farmed Atlantic Salmon Adapted to Different Environments Across Hemispheres

*María Eugenia López[1,2], Tyler Linderoth[3†], Ashie Norris[4], Jean Paul Lhorente[5], Roberto Neira[6] and José Manuel Yáñez[1,7]**

[1] Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile, Santiago, Chile, [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden, [3] Department of Integrative Biology, University of California, Berkeley, CA, United States, [4] Marine Harvest, Kindrum, Fanad, C. Donegal, Ireland, [5] Benchmark Genetics Chile, Puerto Montt, Chile, [6] Facultad de Ciencias Agronómicas, Universidad de Chile, Santiago, Chile, [7] Núcleo Milenio INVASAL, Concepción, Chile

OPEN ACCESS

Edited by:
Maria Saura,
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Spain

Reviewed by:
Roger Vallejo,
Cool and Cold Water Aquaculture Research (USDA-ARS), United States
Andrés Pérez-Figueroa,
University of Porto, Portugal

***Correspondence:**
José Manuel Yañez
jmayanez@uchile.cl

†Present address:
Department of Genetics, University of Cambridge, Cambridge, United Kingdom

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 19 January 2019
Accepted: 26 August 2019
Published: 01 October 2019

Citation:
López ME, Linderoth T, Norris A, Lhorente JP, Neira R and Yáñez JM (2019) Multiple Selection Signatures in Farmed Atlantic Salmon Adapted to Different Environments Across Hemispheres. Front. Genet. 10:901. doi: 10.3389/fgene.2019.00901

Domestication of Atlantic salmon started approximately 40 years ago, using artificial selection through genetic improvement programs. Selection is likely to have imposed distinctive signatures on the salmon genome, which are often characterized by high genetic differentiation across population and/or reduction in genetic diversity in regions associated to traits under selection. The identification of such selection signatures may give insights into the candidate genomic regions of biological and commercial interest. Here, we used three complementary statistics to detect selection signatures, two haplotype-based (iHS and XP-EHH), and one F_{ST}-based method (BayeScan) among four populations of Atlantic salmon with a common genetic origin. Several regions were identified for these techniques that harbored genes, such as *kind1* and *chp2*, which have been associated with growth-related traits or the *kcnb2* gene related to immune system in Atlantic salmon, making them particularly relevant in the context of aquaculture. Our results provide candidate genes to inform the evolutionary and biological mechanisms controlling complex selected traits in Atlantic salmon.

Keywords: selection signatures, *Salmo salar*, Domestication, SNP data, artificial selection

BACKGROUND

Domestication is a complex evolutionary process whereby wild animals or plant populations adapt to environmental conditions created by humans and so involves genetic and developmental changes over multiple generations (Price, 1984; Liu et al., 2017). Since the beginning of domestication, humans have exploited the genetic diversity of various species to model them according to their needs (Driscoll et al., 2009). This has been amplified since the establishment of explicit genetic improvement objectives. As a result of intense selection pressure, dramatic phenotypic changes (Rubin et al., 2012) and substantial and continued genetic improvement have been made in domestic populations over the past decades (Hill and Bunger, 2004).

Domestication in most fish is relatively recent compared with terrestrial animals (Teletchea and Fontaine, 2014; López et al., 2015), but has expanded rapidly over the last decades (Lorenzen et al., 2012), and several breeding programs have been implemented in different aquatic species, such as

tilapia (*Oreochromis niloticus* L), rainbow trout (*Oncorhynchus mykiss* W), coho salmon (*Oncorhynchus kisutch* W), and Atlantic salmon (*Salmo salar* L) among others (Gjedrem, 2010; Gjedrem, 2012; Yáñez et al., 2014). The latter has become one of the most important aquaculture species (FAO, 2016), since it was first farmed in Norway during the 1960s. Despite a generation interval of 3 to 4 years, breeding programs have achieved rapid improvement of economically important traits, such as growth, sexual maturation, and disease resistance (Gjedrem et al., 2012). Domestication and subsequent artificial selection have produced stark phenotypic changes in farmed Atlantic salmon populations (Glover et al., 2017), as evidenced by differences in traits, such as growth and predator awareness, between wild and farmed populations (Thodesen et al., 1999; Glover et al., 2009; Solberg et al., 2012) (Einum and Fleming, 1997).

Positive selection pressures (natural and artificial) experienced by population undergoing selection will cause the frequency of alleles underlying favorable traits to increase rapidly. Linkage disequilibrium (LD) between favorable mutations and neighboring loci will increase and spread, given that there is little opportunity for recombination over the brief time since the onset of intense selection (Sabeti et al., 2002). Analyses of these selection signatures in domestic animals can provide further insights into the genetic basis of adaptation to diverse environments and genotype/phenotype relationships (Oleksyk et al., 2010; Andersson, 2012). Access to genomic data through next-generation sequencing and high-throughput genotyping technologies have made the comparison of genomic patterns of single nucleotide polymorphism (SNP) variation between different livestock breeds possible, allowing for the identification of putative genomic regions and genes under selection in several terrestrial domestic species, including cattle (e.g., Taye et al., 2017), horses (e.g., Avila et al., 2018), sheep (e.g., Ruiz-Larrañaga et al., 2018), and pigs (e.g., Gurgul et al., 2018).

There are several approaches for detecting genomic selection signatures, one of which relies on the length or variability of haplotypes. Directional selection acting on a new, beneficial mutation causes the haplotype harboring the mutation to increase in frequency and to be longer than average. To exploit this pattern for detecting positive selection, Sabeti et al. (2002) proposed the extended haplotype homozygosity (EHH) statistic, which is specifically the probability that two randomly selected haplotypes are identical-by-descent over their entire length around a core SNP (Sabeti et al., 2002). This concept forms the basis for other haplotype homozygosity-based metrics, such as the relative EHH (REHH) (Sabeti et al., 2002) and the widely used integrated haplotype score (iHS) (Voight et al., 2006). iHS compares EHH between derived and ancestral alleles within a population and has the most power to detect selection when the selected allele is at intermediate frequencies in the population (Sabeti et al., 2006; Voight et al., 2006). To detect selection signatures between populations, the cross-population extended haplotype homozygosity test (XP-EHH) compares the integrated EHH profiles between the two populations in the same SNP. This test was designed to detect ongoing or nearly complete selective sweeps in one population (Sabeti et al., 2007). An alternative approach for identifying selection signatures when there are multiple populations for comparison is divergence-based methods, which focus on identifying outlier loci with either higher or lower allele frequency differences among populations than expected without selection (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008; Excoffier et al., 2009). One common approach for quantifying the degree of genetic differentiation between populations is through the fixation index, F_{ST}, (Wright, 1951). An unusually high F_{ST} value at a given locus can be indicative of directional selection. Divergence approaches to identify signals of selection have been successful in several domestic species including swine (Cesconeto et al., 2017), sheep (Manunza et al., 2016), and cattle (Maiorano et al., 2018) among others.

Although previous studies have already been carried out to detect selection signatures in Atlantic salmon (Mäkinen et al., 2014; Gutierrez et al., 2016; Liu et al., 2017; López et al., 2018), using multiple different strains adapted to different culture conditions across hemispheres, to explore how genetic variation among them differs, has not been done yet. Herein, we used an Affymetrix 200K SNP array data set to investigate selection signatures in farmed Atlantic salmon populations from the same origin, and subsequently cultivated in Ireland and Chile. We found evidence of selection using two haplotype-based approaches iHS and XP-EHH and one F_{ST}-based method, BayeScan, in the genomes of four Atlantic salmon populations. These findings are important because they highlight regions of the genome that might benefit economically relevant attributes, such as growth, resistance to local diseases, and adaptation to specific environmental conditions.

MATERIALS AND METHODS

Samples, Genotyping, and Quality Control

This study was performed using a total of 270 individuals from four populations (Pop-A, n = 40; Pop-B, n = 71; Pop-C, n = 85; Pop-D, n = 74) derived from the Mowi strain. This strain comes from one of the first farmed Atlantic salmon populations, which was established with fish from west coast rivers in Norway, with major contributions from River Bolstad in the Vosso watercourse, River Årøy, and possibly from the Maurangerfjord area (Verspoor et al., 2007). Salmon from the Vosso and Årøy rivers are characterized by large size and late maturity (Verspoor et al., 2007). Phenotypic selection for growth, late maturation and fillet quality was the focus in this population until 1999 (Glover et al., 2009). Ova from this population were imported into the Fanad Peninsula, Ireland, between 1982 and 1986 to establish an Irish-farmed population (Norris et al., 1999). Individuals from this population comprise Pop-A, which we estimate had been under artificial selection for growth for at least 10 generations prior to sampling. Similarly, ova from this farmed, Irish population were introduced into Chile in the early 1990s to establish separate farmed populations in the Los Lagos Region (42°S 72°O) and the Magallanes Region (53°S 70°O). Pop-B and Pop-C correspond to samples from two different populations in the Los Lagos Region that were initially founded with fish from different year-classes. Samples from Pop-D represent one population founded

in the Magallanes Region. The three Chilean populations were subsequently adapted to the biotic and abiotic conditions present in southern hemisphere. These populations experienced four generations of selective breeding for growth in Chilean farming conditions prior to sampling, which occurred at the same time that Pop-A was sampled in 2014.

All populations were genotyped using Affymetrix's Atlantic salmon 200K SNP Chip described in Yáñez et al. (2016). We performed SNP quality control using the Axiom Genotyping Console (GTC, Affymetrix) and SNPolisher (an R package developed by Affymetrix), which i) removed SNPs that did not conform high-quality clustering patterns as outlined by Affymetrix, ii) removed SNPs with genotype call rate lower than 95%, and iii) discarded individuals with genotyping call rate under 90%. As part of the validation of the SNPs chip used in this study, Yáñez et al. (2016) identified loci significantly deviating from Hardy–Weinberg equilibrium in eight populations separately and removed these sites if they were deviating from Hardy–Weinberg equilibrium among all populations. In addition, we limited our analyses to SNPs that mapped to chromosomes in the newest version of the Atlantic salmon reference genome, ICSAG_v2 (GenBank: GCA_000233375.4), which comprised 149,060 SNPs.

Genetic Diversity, LD, and Population Structure

We evaluated genetic diversity in terms of the observed heterozygosity (H_O) and expected heterozygosity (H_E) calculated with PLINK v1.09 (Purcell et al., 2007). We calculated the pair-wise LD as the Pearson's squared correlation coefficient (r^2) for each population and within chromosomes using PLINK v1.09 (Purcell et al., 2007). For each SNP pair, bins of 100 kb were created based pairwise distance. To investigate population structure, we performed a principal component analysis (PCA) based on genotypes as implemented in PLINK v1.09 and inferred individual ancestry proportions with ADMIXTURE 1.2.2 (Alexander et al., 2009). For the admixture analysis, we performed 200 bootstraps with a number of ancestral lineages (K) ranching from 1 to 20. Ten-fold cross validation (CV = 10) was specified, and we retained results from the K having the lowest cross-validation error. The aforementioned analyses were conducted using a total of 21,950 SNPs, which had a minor allele frequency (MAF) larger than 0.05, were in Hardy–Weinberg equilibrium, and which had LD values of at most 0.4 (to minimize possible confounding effects of LD on the patterns of genetic structure).

Selection Signatures, Gene Annotation, and Functional Analyses

To identify genomic regions harboring selection signatures, we used one within population iHS and two between-population methods (XP-EHH and BayeScan) over a subset of 120,316 SNPs that had MAF > 0.05 among all populations.

(1) iHS. The iHS score for detecting selection is based on the ratio of EHH for haplotypes anchored with the ancestral versus derived allele. The ancestral allele state for our Atlantic salmon populations is unknown and so to avoid losing SNPs by trying to polarize them from publicly available outgroup references, we assumed that the major allele represented the ancestral state as in Bahbahani et al. (2015). We phased the haplotypes using Beagle v.5.0 (Browning and Browning, 2009). Single-site iHS values across the genome were calculated for each populations using the REHH package (Gautier and Vitalis, 2012). These per site iHS values were standardized so that they were approximately distributed according to a standard normal distribution. We required candidate-selected regions to have at least two SNPs ≤ 500 kb apart, each with iHS scores with $-\log_{10}(p \text{ value})$ of at least three (p value ≤ 0.001) based on a one-tailed test assuming that the standardized iHS ~ N(0,1).

(2) XP-EHH. The XP-EHH statistic compares the integrated EHH between two populations at the same SNP, to identify selection based on overrepresented haplotypes in one of the populations (Sabeti et al., 2007). We evaluated three different pairs of populations with this method Pop-B/Pop-A, Pop-C/Pop-A, and Pop-D/Pop-A. This design was used because of the main objective of this study was to assess how selective pressures have affected populations cultivated in Chile, relative to their founding population, Pop-A, which was used as the reference population. Therefore, we excluded the comparisons between Chilean populations. The XP-EHH statistics were calculated as $\ln(I_{PopO}/I_{PopR})$, where I_{PopO} is the integrated EHH for the observed populations and I_{PopR} is the integrated EHH value of the reference population. Negative XP-EHH scores suggest selection in the "reference" population, whereas positive scores suggest selection acting in the "observed" population. A $-\log_{10}(p \text{ value})$ of three (p value ≤ 0.001) was used as the lower threshold for considering XP-EHH score as significant evidence of selection and at least two SNPs ≤ 500 kb apart.

(3) BayeScan. We used the Bayesian likelihood method implemented in BayeSCAN v.2.1 to estimate the posterior probability that loci are experiencing selection (Foll and Gaggiotti 2008). This method models allele frequencies in subpopulations derived from a single ancestral population using Dirichlet distributions, which allows for estimating the degree of coancestry within each of these subpopulations through the sum of population-specific, β, and locus-specific, α, effects, making outlier detection robust to confounding complex demographic histories. By estimating the posterior probabilities for both the model including both effects and the model omitting the locus-specific effect, the posterior probability (and posterior odds) for selection at a specific locus can be obtained. When $\alpha > 0$ for a specific locus, it is evidence of directional selection acting on that locus, whereas $\alpha < 0$ suggests balancing or purifying selection. This method was run with 5,000 burn-in iterations, followed by 10,000 iterations with a thinning interval of 10. We evaluated the same three pairs of populations of XP-EHH method: Pop-B/Pop-A, Pop-C/Pop-A, and Pop-D/Pop-A. We considered candidate loci under selection as those having a Bayes factor of at least 32 ($-\log_{10} = 1.5$) and a positive value of α (directional selection), corresponding to a posterior probability of 0.97 and considered as being "very strong" evidence of selection and as in iHS and XP-EHH, we required the candidate selected regions to have at least two SNPs ≤500 kb apart.

Gene Functional Annotation

Genomic regions harboring SNPs showing evidence of selection were annotated based on the ICSAG_v2 reference genome (Lien et al., 2016). We defined the position of the first and last SNP as boundaries of regions putatively under selection using BedTools (Quinlan and Hall, 2010). Gene transcripts from these candidate regions were aligned (using blastx) (Altschul et al., 1990) to the zebra fish (*Danio rerio*) peptide reference database (downloaded from http://www.ensembl.org/) to determine gene identify. As evidence of homology, we used an e-value $\simeq 0$ and then retrieved the zebra fish gene identifiers information from the ensemble biomart database (http://www.ensembl.org/index.html). Functional annotation of detected genes was performed using DAVID (Huang et al., 2009) with gene list of zebra fish (*Danio rerio*) as reference in Gene Ontology (GO) analysis.

RESULTS

Genetic Diversity and Structure

We performed PCA based on genotypes to look at the genetic relationship among individuals in our sample. The first and second components accounted for 14.2% and 10.3% of the genetic variation, respectively (**Figure 1**). Pop-A and Pop-C showed close genetic relationship to each other and were most distant to Pop-D from the Magallanes Region along PC1. Pop-B lies between the Pop-A/Pop-C cluster and Pop-D along PC1, with some overlap with Pop-C, which was introduced into the same Los Lagos Region as Pop-B. Overall, principal components showed low genetic variation between populations, but higher within populations, especially in Pop-D that exhibits the most difference among individuals along PC1. Also noteworthy is that Pop-D, with the highest observed heterozygosity (**Table 1**), is uniformly farther to the other farmed populations, except for some individuals from Pop-B. We also performed an Admixture analysis to determine the composition of ancestral lineages among individuals. We found that 11 ancestral lineages were optimal for describing the ancestry of the individuals across the four populations (**Figure 2**). Consistent with the PCA and having the lowest heterozygosity, Pop-A individuals are all relatively the most similar among the populations in terms of their ancestral proportions, being dominated by one ancestral lineage. In contrast, Pop-D individuals tend to be dominated by a single ancestral lineage, but among individuals, the represented lineages are quite different, which is consistent with Pop-D individuals being quite different from each other in the PCA. Pop-B and Pop-C show similar degrees of mixed ancestry, though the dominant lineage is different between the two.

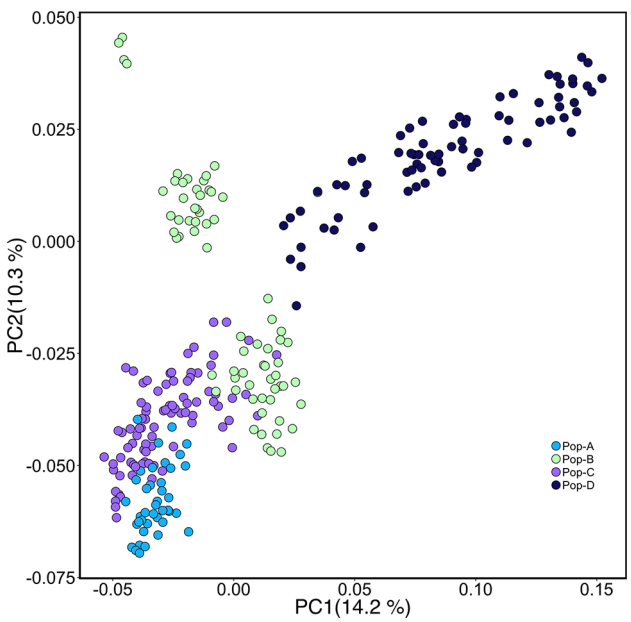


FIGURE 1 | Principal components analysis (PCA) of genetic differentiation among individuals. Each point represents one individual, and different colors represent populations.

Observed heterozygosity levels were similar across the four domestic populations and were slightly higher than expected for populations A, B, and C, and even more so for population D. All these genetic diversity measures were statistically significant ($p < 0.05$, Kruskal–Wallis test) (see **Table 1**). Overall LD results revealed similar patterns for Pop-A and Pop-D, which presented longer range of LD and slower decay in comparison with Pop-B and Pop-C, that also presented similarity between them and a substantial faster LD decay (**Figure 3**). LD measures (r^2) of each chromosome and population are shown in **Table S1** and **Figure S1**. Similar patterns were observed when the chromosomes were analyzed separately. Nevertheless, LD decay in Pop-A was noticeably stronger in chromosomes 2, 9, 19, and 29, whereas LD decay in Pop-D was stronger in chromosomes 13, 17, and 26 (**Figure S1**).

Candidate Regions Under Selection—iHS

We looked for evidence of selection by comparing the decay of association between alleles from the major versus minor allele at core SNPs using iHS. We found 115, 63, 142, and 467 core SNPs with significant iHS statistics ($p \leq 0.001$) for Pop-A, -B, -C, and -D respectively (**Figure 4**, **Table 2**). We find 27, 12, 23, and 83 regions in these respective populations with at least two significant SNPs that are ≤ 500 kb apart, which we classify as putatively, selected regions.

Candidate regions for Pop-A were on Ssa01, Ssa05, and Ssa22. The candidate regions having SNPs with the most significant

TABLE 1 | Genetic diversity values in terms of Observed heterozygosity (H_o) and Expected heterozygosity (H_E) across four Atlantic salmon populations used in this study.

Population	H_o	H_e
Pop-A	0.4 ± 0.13	0.39 ± 0.11
Pop-B	0.41 ± 0.11	0.41 ± 0.1
Pop-C	0.41 ± 0.11	0.41 ± 0.1
Pop-D	0.47 ± 0.17	0.39 ± 0.11

All these genetic diversity measures were statistically significant ($p < 0.05$, Kruskal–Wallis test).

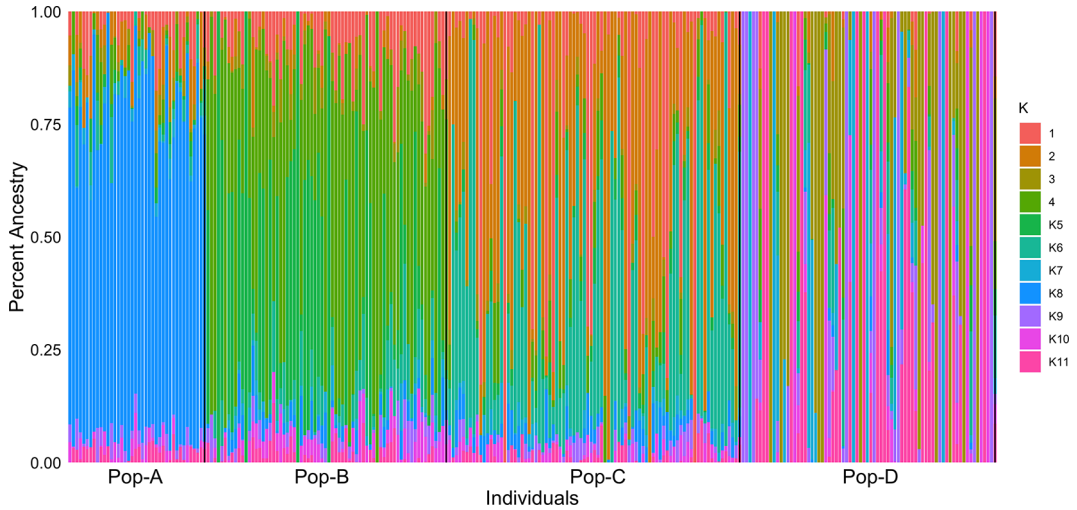


FIGURE 2 | Individual assignment probabilities generated with ADMIXTURE (1⩽K⩽11). Each color represents a cluster, and the ratio of vertical lines is proportional to assignment probability of and individual to each cluster.

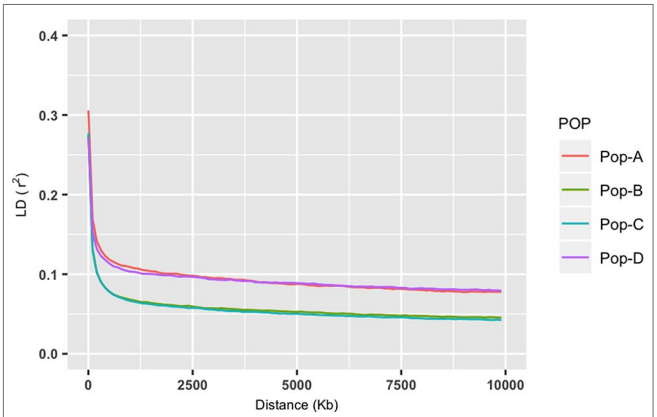


FIGURE 3 | Decay of average linkage disequilibrium (r^2) over distance across the four farmed populations. Different color lines represent populations: Pop-A = Red, Pop-B = Green; Pop-C = Turquoise and Pop-D = Purple.

iHS scores were on Ssa05, Ssa10, and Ssa14, which contained the genes *igfbpl1* and *mipol1*.

Pop-B had 12 regions with an average length of ~ 250 kb putatively under selection distributed among five chromosomes. The highest iHS score was for a region found on Ssa13 [-log(*p* value) = 4.17] containing 26 genes including the *soga1* gene. Pop-C had 23 candidate regions that were on average ~370 kb long, and which spanned a total of 165 genes. The 1,570-kb-long region with one of the most significant iHS score was on Ssa22, and spanned the genes *kcnkf*, *sc61a,* and *mstn1*. Pop-D had the most significant number of SNPs (467) and had 83 putatively selected genomic regions under our criteria. Most of these regions were located on Ssa01, Ssa10, Ssa13, and Ssa26 and spanned genes, such as *haus2*, *itfg*1, and *phkb*. Details of the total regions and genes can be found in **Supplementary Tables S2** and **S5**, respectively.

Candidate Regions Under Selection—*XP-EHH*

We compared the decay of LD from a core SNP as measured by EHH between the Norwegian source population and the three derived Chilean populations (Pop-B/Pop-A, Pop-C/Pop-A, Pop-D/Pop-A) to detect regions having unusually high EHH and overrepresented haplotypes consistent with selection. In total, we detected 482 (Pop-B/Pop-A), 800 (Pop-C/Pop-A), and 207 (Pop-D/Pop-A) XP-EHH outlier SNPs indicative of selection (**Figure 5**, **Table 3**). The sign of the XP-EHH score indicates which population selection is acting on. Here, negative scores suggest selection in Pop-A. Most significant SNPs, which we considered as those with XP-EHH score $p \leq 0.001$, had negative scores, suggestive of selection in the Irish source population. The Pop-C/Pop-A and Pop-D/Pop-A comparisons yielded 38 and 3 significant SNPs with positive scores respectively, suggesting that the C and D populations underwent selection after their introduction into Chile. The significant, positive scores suggesting selection in Pop-C were found on Ssa16 within two regions spanning a total of 664.2 kb and which harbored 17 genes. The significant SNPs pointing to selection in Pop-D were located on Ssa14 in an 18.4-kb region, which contained the gene *agla*. XP-EHH did not detect selection signatures in Pop-B, as all significant scores for the Pop-B/Pop-A pair were negative. We classified potential genomic regions under selection as those containing two or more significant, adjacent SNPs less than 500 kb apart. After merging overlapping regions, we identified 34, 28, and 23 candidate regions from the Pop-B/Pop-A, Pop-C/Pop-A, and Pop-D/Pop-A comparisons respectively, which were all suggestive of selection in Pop-A. The average lengths of the candidate regions are approximately 338 kb for Pop-B/Pop-A, 546.5 kb for Pop-C/Pop-A, and 139 kb for Pop-D/Pop-A. Together, these regions span a total of 667 genes. Details of the total regions and genes detected by XP-EHH can be found in **Supplementary Table S3** and **S6**, respectively.

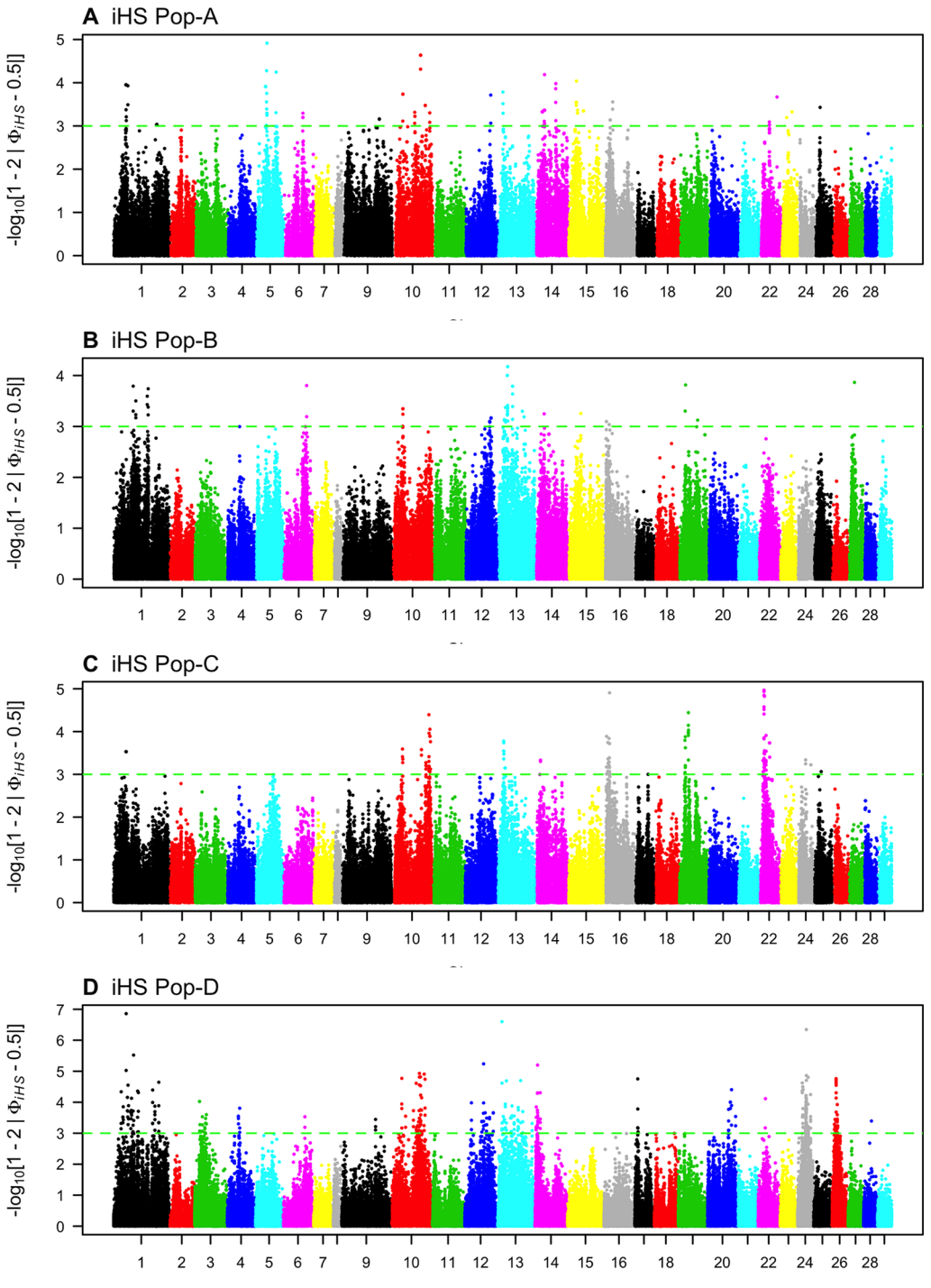


FIGURE 4 | Genome-wide distribution of -$\log_{10}$(*P* value) of standardized integrated haplotype score (iHS) across four Atlantic salmon populations: **(A)** Pop-A, **(B)** Pop-B, **(C)** Pop-C, and **(D)** Pop-D.

Candidate Regions Under Selection—*BayeScan*

We used the Bayesian approach for estimating the posterior odds of selection acting at particular loci based on pairwise divergence between ancestral and derived populations implemented in BayeScan. By applying the BayeScan method to Pop-B/Pop-A, Pop-C/Pop-A, and Pop-D/Pop-A population pairs we, respectively, found 167, 155, and 193 SNPs with posterior odds ratios above 32, which was our threshold for showing significant evidence of selection (**Figure 6**, **Table 4**). F_{ST}-based methods do not directly

TABLE 2 | Ten genome regions spanning the strongest detected selection signatures by iHS in each population.

POP	CHR	START	END	-log(*p*)	|iHS|	N SNPs	SIZE (kb)
Pop-A	1	35662318	35684677	3.9515	3.8634	4	22.4
	1	40728165	40728699	3.9306	3.8516	2	0.5
	5	25918328	25932901	3.9169	3.8439	2	14.6
	5	28372137	29065939	4.2779	4.0432	5	693.8
	5	29574408	29842752	4.9158	4.3751	4	268.3
	5	55278111	55732536	4.2454	4.0256	2	454.4
	10	79382450	79401333	4.6374	4.2331	3	18.9
	14	24674586	25715785	4.188	3.9944	6	1041.2
	14	56736246	57120611	3.98	3.8794	4	384.4
	15	22773166	23073140	4.0388	3.9122	6	300
Pop-B	1	55995699	56003301	3.7914	3.7725	2	7.6
	1	63381907	63519535	3.5006	3.602	3	137.6
	1	95895287	95964374	3.5929	3.6568	2	69.1
	1	98490168	98568189	3.7394	3.7425	3	78
	6	65448844	65496448	3.8025	3.7788	2	47.6
	10	29948442	30759289	3.347	3.509	4	810.8
	12	71810541	71833088	3.1679	3.3978	2	22.5
	13	22110373	22139660	3.371	3.5237	3	29.3
	13	27127510	28267153	4.1737	3.9866	10	1139.6
	13	41965178	42139618	3.8649	3.8144	15	174.4
Pop-C	10	104686655	105233083	4.3935	4.1051	10	546.4
	10	107544485	107633657	4.0535	3.9204	4	89.2
	16	6030690	6249119	3.8867	3.8268	3	218.4
	16	12985501	13367987	3.8395	3.7999	4	382.5
	16	14071951	14921512	4.9062	4.3703	7	849.6
	19	16762570	17099010	3.7932	3.7735	3	336.4
	19	17814180	18048311	3.877	3.8213	4	234.1
	19	26510295	26774029	4.4423	4.131	7	263.7
	22	16108465	17678398	4.9676	4.401	23	1569.9
	22	21553029	22158359	3.9096	3.8398	6	605.3
Pop-D	1	36254384	36351335	6.8581	5.267	2	97
	1	57451430	57811304	5.5208	4.6699	2	359.9
	10	80280055	81084212	4.9294	4.3819	6	804.2
	10	83621134	84353833	4.8176	4.3255	9	732.7
	10	93572235	93962188	4.9072	4.3708	3	390
	12	53402445	54250457	5.2386	4.5345	6	848
	13	14998846	15196522	6.6012	5.1573	5	197.7
	14	8208052	9149527	5.1987	4.5151	10	941.5
	24	24577538	26845034	6.3462	5.0462	54	2267.5

indicate in which population selection is acting; therefore, we describe our findings in terms of the population pairs. Since we expect regions that are truly under selection to have clusters of highly diverged SNPs in LD, we considered only regions containing at least two significant SNPs that were less than 500 kb adjacent to each other as being strong selection candidates. Under this criterion 104, 98, and 121 SNPs with posterior odds ratios of selection above 32 remain of interest for the Pop-B/Pop-A, Pop-C/Pop-A, and Pop-D/Pop-A comparisons, respectively. Clusters of SNPs identified as being in or adjacent to putatively selected regions from the Pop-B/Pop-A comparison represent 31 regions that are, on average, ~96.8 kb long and which harbored 58 genes. The Pop-C/Pop-A comparison showed 98 highly diverged regions among 29 regions that were, on average, ~220.7 kb long and which spanned 200 genes. Finally, the Pop-D/Pop-A comparison revealed 28 candidate regions that were, on average, ~153.6 kb long and contained 130 genes. Only two SNPs among these candidate regions showed evidence of selection among the three population pairs, which were located on Ssa29 in association with the *kmt2ca* gene. Twenty SNPs suggestive of selection were shared between Pop-B/Pop-A and Pop-C/Pop-A and were associated to regions that harbor 12 genes *rabac1, znf1030, tpi1b, si:ch211-206a7.2, znf1041, lpcat3, atp1a3b, zgc:158654,* and *myh10* on Ssa02, *znf385d* on Ssa05, *agbl4* on Ssa10, and CR388166.1, and *kmt2ca* on Ssa29. Four candidate SNPs were common to Pop-C/Pop-A and Pop-D/Pop-A and two between Pop-B/Pop-A and Pop-D/Pop-A, which correspond to the *kmt2ca* gene shared among three population pairs. Details of the total regions and genes detected by BayeScan can be found in **Supplementary Tables S4** and **S7**, respectively.

Gene Ontology for Candidate Genes Under Selection

To further explore the functions of the candidate genes spanned by regions showing evidence of selection from the iHS, XP-EHH, and BayeScan analyses, we annotated the candidate genes using the DAVID browser (https://david-d.ncifcrf.gov). The candidate genes were enriched in 37 gene ontology (GO) terms overall, most of them

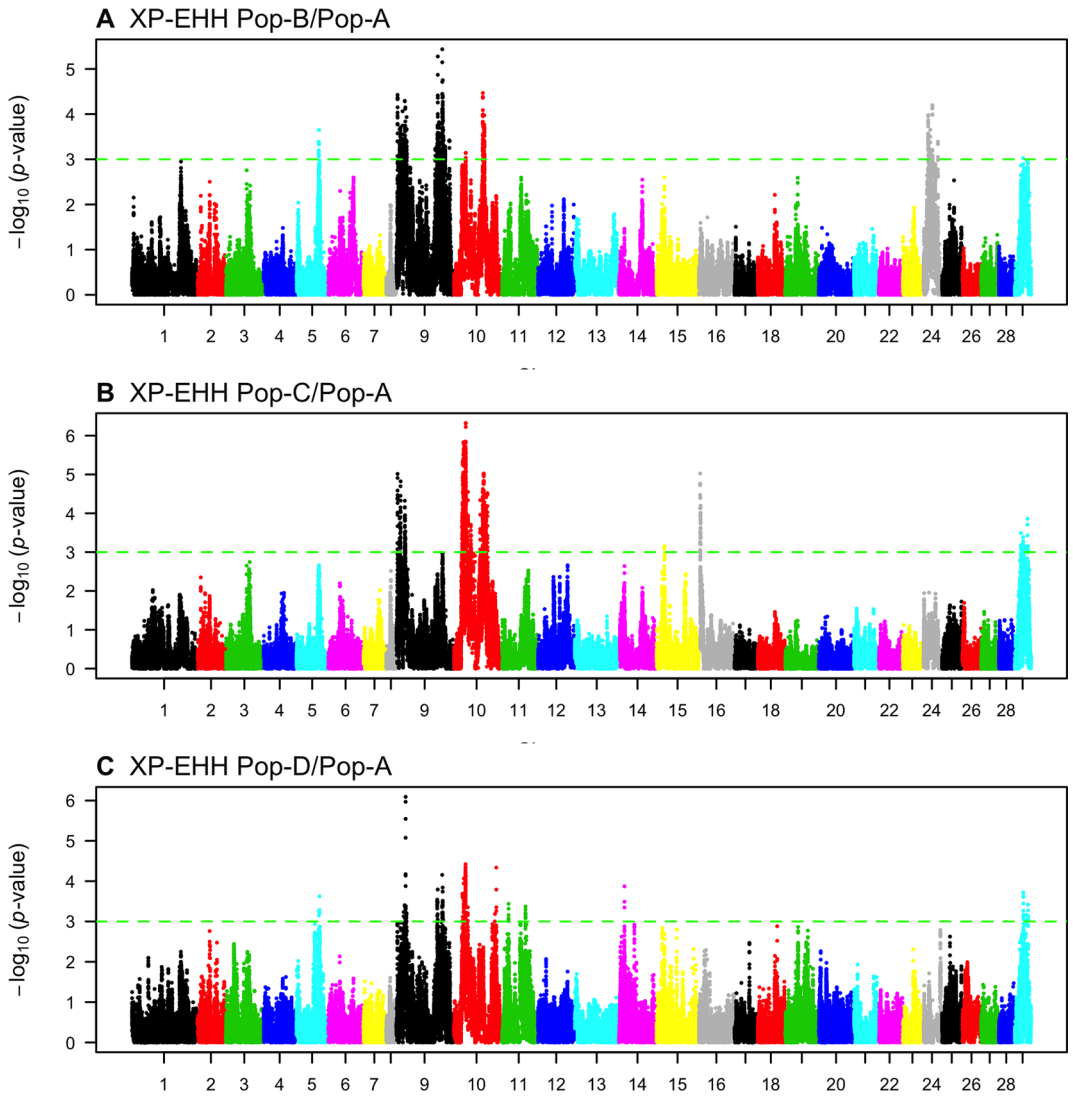


FIGURE 5 | Genome-wide distribution of -$\log_{10}$(*P* value) of standardized cross-population extended haplotype homozygosity (XP-EHH) scores across three pairwise Atlantic salmon populations: **(A)** Pop-B/Pop-A, **(B)** Pop-C/Pop-A, and **(C)** Pop-D/Pop-A.

TABLE 3 | Genome regions spanning the strongest detected selection signatures by XP-EHH in populations A, C and D.

POP	CHR	START	END	-log(*P*)	XP-EHH	N SNPs	Size (kB)
Pop-A	10	28741972	31140475	6.324	-5.0365	153	2398.5
	9	23728364	24144245	6.0912	-4.9328	15	415.9
	10	24160722	26099914	5.8281	-4.8132	127	1939.2
	9	113910288	114187655	5.4365	-4.6298	30	277.4
	10	21739331	23180890	5.3423	-4.5847	54	1441.6
	9	101786257	103293781	5.2766	-4.553	56	1507.5
	10	73472292	74738689	5.021	-4.4276	73	1266.4
	9	3674860	4026195	5.0148	-4.4245	14	351.3
	9	11161334	11559014	4.823	-4.3282	19	397.7
	9	114997862	115904242	4.7511	-4.2916	44	906.4
Pop-C	16	3564058	3808523	5.0253	3.2964	13	244.5
	16	4345514	4765204	4.4036	3.2915	25	419.7
Pop-D	14	14636389	14654809	3.872	3.5093	3	18.4

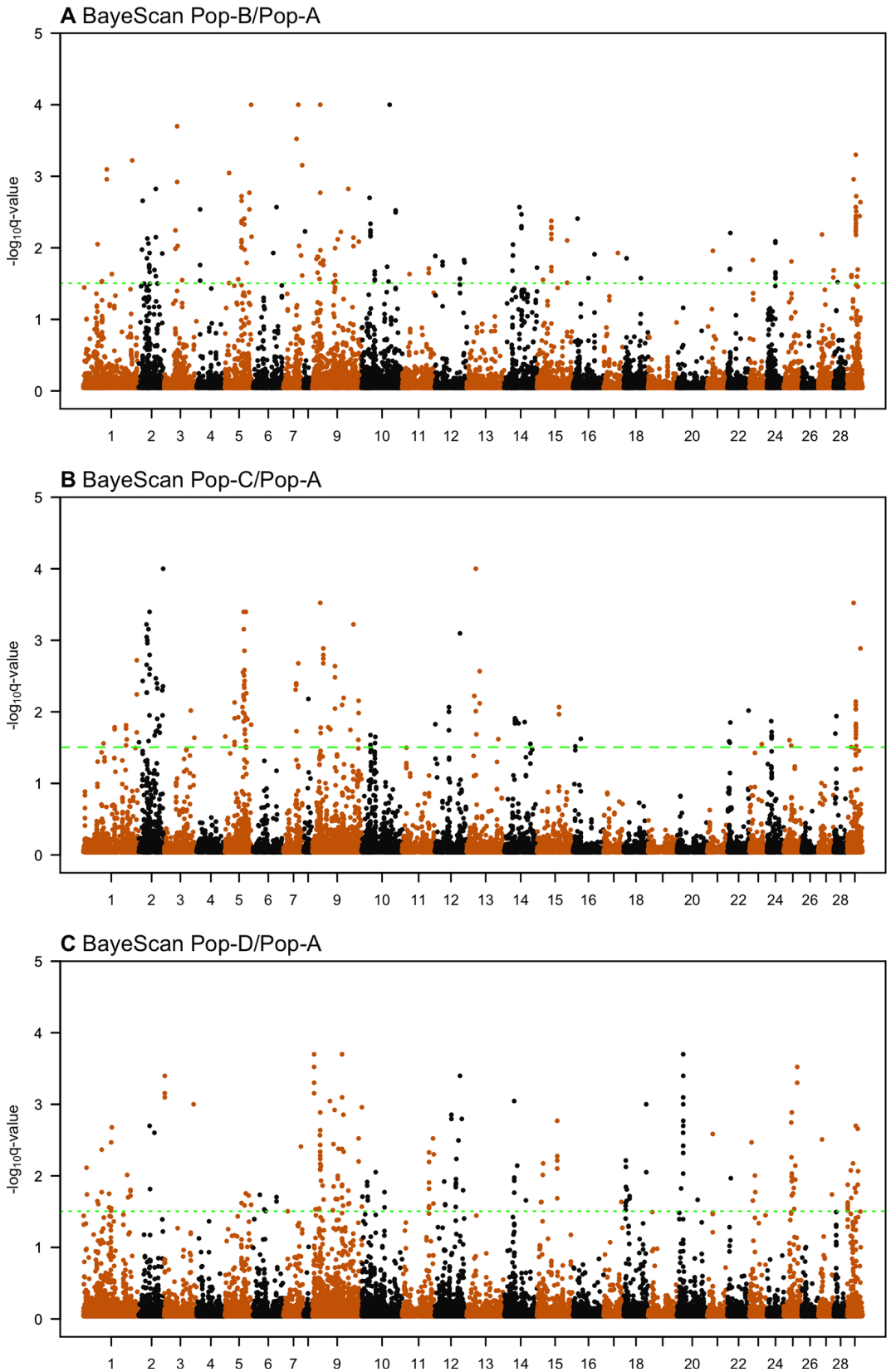


FIGURE 6 | Genome-wide distribution of -log10(*q* value) in BayeScan analysis across three pairwise Atlantic salmon populations: **(A)** Pop-B/Pop-A, **(B)** Pop-C/Pop-A, and **(C)** Pop-D/Pop-A.

TABLE 4 | Ten genome regions spanning the strongest detected selection signatures by BayeScan method in each population pair.

POP	CHR	START	END	-log(*q* value)	N SNPs	SIZE (kb)
Pop-B/Pop-A	1	66642439	66648870	3.097	2	6.4
	2	48416445	48567681	2.824	3	151.2
	3	37023703	37052183	3.699	2	28.5
	5	11553116	11556394	3.046	2	3.3
	5	47250892	47494177	2.721	8	243.3
	5	69864532	69865664	2.770	2	1.1
	7	53824902	53839042	3.155	2	14.1
	9	22192894	22527645	4.000	4	334.8
	29	23820153	24379023	3.301	10	558.9
	29	25107616	25137079	2.721	6	29.5
Pop-C/Pop-A	2	21743330	22285719	3.222	4	542.4
	2	24203593	24203644	3.000	2	0.1
	2	27316859	27731651	3.155	2	414.8
	2	30394158	31352454	3.398	5	958.3
	2	69206072	69622942	4.000	2	416.9
	5	52915411	53613743	3.398	5	698.3
	5	59616884	59678559	3.398	4	61.7
	9	30961027	30994613	2.886	4	33.6
	13	25691366	25715347	4.000	2	24.0
	29	23852604	24289616	2.886	12	437.0
Pop-D/Pop-A	3	1316421	1317893	3.398	3	1.5
	9	4536926	4590569	3.699	4	53.6
	9	22080850	22146356	2.886	9	65.5
	9	84229138	85051608	3.699	6	822.5
	9	141700047	141700106	2.959	2	0.1
	14	28094905	28343120	3.046	4	248.2
	18	64503159	64648197	3.000	2	145.0
	20	17158525	17477234	3.699	10	318.7
	25	22716335	22760611	2.886	6	44.3
	25	38351034	38355710	3.523	2	4.7

population specific (**Table 5**). Four GO categories were common between Pop-A and Pop-B (single-multicellular organism process, single-organism developmental process, regulation of metabolic process, and anatomical structure development) and one between Pop-C and Pop-D (animal organ development). The remaining GO categories were unique to each population.

DISCUSSION

In this study, we used three complementary tests to detect selection signatures within and between four Atlantic salmon populations with Norwegian origin. We used the iHS test to scan for selection signatures within populations and XP-EHH and BayeScan to find evidence of selection in terms of divergence of the Chilean populations to their ancestral Irish population. We detected several genomic regions under putative selection across all of the populations evaluated, which provides insight into the genes contributing to traits of importance to Atlantic salmon farming. It is important to mention that these findings should be interpreted with caution since other evolutionary and demographic process, such as bottlenecks and differences in the amount of genetic drift resulting from different effective populations sizes, can produce patterns of genetic diversity that mimic selection leading to the finding of possible false positives as well. However, the selection detection methods we used have all been shown to be robust to these confounding effects.

Structure and Diversity

To examine genetic population structure and relationships among the major groups of salmon, we conducted an ADMIXTURE analyses based on high-quality SNP data. This analysis revealed that 12 ancestral lineages contribute to the modern gene pool represented by the four farmed populations, which was expected considering the admixed origin of these populations (Verspoor et al., 2007). The four populations used in this study are derived from the Mowi strain, which was created using samples from several rivers along the west coast of Norway (Norris et al., 1999). The population with the lowest level of admixture was Pop-A, which was also the population with the lowest genetic diversity, a condition that could reflect a better culture management, as well as intense artificial selection that erodes genetic variation through mating related individuals (Gjedrem, 2005). Pop-B and Pop-C which were introduced into the same region in Chile have very similar amounts of heterozygosity and similar degrees of admixture though the dominant lineages are different, which was expected due to the similar breeding practices and environmental conditions to which they have been subjected. Pop-D, however, showed the highest level of heterozygosity and a more complex pattern of admixture, whereby a single ancestral lineage is highly represented within individuals but with many ancestral lineages present among individuals. This pattern may, in part, reflect lower artificial selection pressure. Recent genetic introgression cannot be ruled out for Pop-D given the potential for crossing with different strains for management reasons. LD analysis revealed that overall

TABLE 5 | Biological processes enriched in genes detected by iHS and XP-EHH in each Atlantic salmon population.

Population	Biological Process	GO Term	%	*p*	Benjamini
Pop-A	Cellular metabolic process	GO:0044237	36.8	3.0E-4	3.7E-2
	Organic substance metabolic process	GO:0071704	38.7	9.4E-4	5.6E-2
	Primary metabolic process	GO:0044238	37.1	1.2E-3	4.8E-2
	Catabolic process	GO:0009056	5.7	2.3E-2	5.1E-1
	Single-multicellular organism process	GO:0044707	19.1	4.7E-2	7.0E-1
	Developmental induction	GO:0031128	0.4	5.9E-2	7.1E-1
	Single-organism developmental process	GO:0044767	19.2	6.7E-2	7.1E-1
	Regulation of metabolic process	GO:0019222	14.0	7.6E-2	7.0E-1
	Anatomical structure development	GO:0048856	19.1	9.7E-2	7.5E-1
Pop-B	Regulation of signaling	GO:0023051	14.5	8.8E-3	4.7E-1
	Regulation of cellular process	GO:0050794	45.2	1.4E-2	4.0E-1
	Regulation of metabolic process	GO:0019222	22.6	3.8E-2	6.0E-1
	Anatomical structure morphogenesis	GO:0009653	17.7	4.8E-2	5.9E-1
	Regulation of response to stimulus	GO:0048583	12.9	5.0E-2	5.2E-1
	Cellular component organization	GO:0016043	24.2	5.1E-2	4.7E-1
	Single-organism developmental process	GO:0044767	27.4	6.2E-2	4.8E-1
	Anatomical structure development	GO:0048856	27.4	6.6E-2	4.6E-1
	Single-multicellular organism process	GO:0044707	25.8	9.8E-2	5.6E-1
	Methylation	GO:0032259	4.8	9.9E-2	5.3E-1
Pop-C	Heart development	GO:0007507	5.3	2.4E-2	1.0E0
	Regulation of cell communication	GO:0010646	8.8	2.7E-2	9.7E-1
	Regulation of signal transduction	GO:0009966	8.2	2.8E-2	9.2E-1
	Animal organ development	GO:0048513	14.7	3.0E-2	8.8E-1
	Organ morphogenesis	GO:0009887	6.5	3.3E-2	8.4E-1
	Digestive tract development	GO:0048565	2.4	3.7E-2	8.2E-1
	Muscle system process	GO:0003012	2.4	4.9E-2	8.6E-1
	Tissue development	GO:0009888	9.4	6.3E-2	8.9E-1
	Cellular developmental process	GO:0048869	13.5	6.5E-2	8.7E-1
	Phosphorus metabolic process	GO:0006793	12.4	8.1E-2	9.0E-1
	System development	GO:0048731	17.6	9.7E-2	9.2E-1
Pop-D	Pancreas development	GO:0031016	1.7	5.1E-3	9.0E-1
	Cellular lipid metabolic process	GO:0044255	4.0	6.3E-3	7.6E-1
	Regulation of blood pressure	GO:0008217	1.0	1.1E-2	8.1E-1
	Lipid metabolic process	GO:0006629	4.7	1.6E-2	8.3E-1
	Gland development	GO:0048732	2.2	1.6E-2	7.7E-1
	Forebrain development	GO:0030900	1.5	2.6E-2	8.6E-1
	Small molecule metabolic process	GO:0044281	6.2	4.4E-2	9.5E-1
	Atrioventricular canal development	GO:0036302	0.5	5.0E-2	9.5E-1
	Organic acid metabolic process	GO:0006082	3.5	5.1E-2	9.3E-1
	Embryonic organ development	GO:0048568	3.5	7.0E-2	9.6E-1
	Animal organ development	GO:0048513	11.4	8.6E-2	9.7E-1
	Single-organism biosynthetic process	GO:0044711	4.2	9.2E-2	9.7E-1

LD decays more rapidly in Pop-B and Pop-C over short physical distances and is lower than Pop-A and Pop-D. The pattern of LD in Pop-A is consistent with its lower heterozygosity level. However, similar pattern was observed in Pop-D, likely due to higher level of admixture in this population, where several ancestral lineages can be observed. Chromosomal LD decay followed similar patterns, but in Pop-A, LD decay was noticeably higher in chromosomes 2, 9, 11, 19, and 29, which is agreed with a greater number of regions detected under selection in those chromosomes. Conversely, in chromosome 26, Pop-D showed the highest value of LD (r^2 = 0.12), probably related to a larger region under selection detected in this population. The results presented here also reinforce the notion that exposure to different management and environmental conditions over just a few generations (at least four in this particular case) is sufficient to generate large changes in the genetic structure of farmed Atlantic salmon populations with the same genetic origin.

Selection Signatures

Pop-D had regions showing the strongest evidence for selection as well as the most candidate regions according to the iHS test. Although the iHS test has a lower power to detect selection under nearly complete sweeps (Sabeti et al., 2007; Simianer et al., 2010), it has greater power when selected alleles are at intermediate frequencies. Pop-D has experienced weaker artificial selection pressure than the other populations used in this study (Jean Paul Lhorente, personal communication), and so the higher number of putatively selected regions identified in this population by iHS may reflect more sweeps at intermediate frequencies because they are taking relatively longer to complete under weaker selection. In addition, this population is located in the Magallanes Region in Chile, which exposes salmon to more extreme environmental conditions than in the Los Lagos region where Pop-B and Pop-C were introduced. Therefore, the selection imposed by the natural

environmental may also contribute to a relatively high number of selected regions in Pop-D. In contrast to iHS, XP-EHH is powerful at detecting complete or nearly complete selective sweeps (Sabeti et al., 2007). According to the XP-EHH method, Pop-A shows the greatest number of regions under selection across the genome, which is consistent with XP-EHH having greater power to identify selection in regions that experienced older selection events (Sabeti et al., 2007; Klimentidis et al., 2011) than iHS since Pop-A is the oldest population in the present study while also being subjected to more intense artificial selection. We identified several putative directional selection targets using BayeScan, but given the nature of F_{ST}-based methods we are unable to directly identify which population in a pairwise comparison is experiencing selection from the posterior odds alone. Low overlap in selected regions identified with haplotype-based and single-SNP F_{ST}-based approaches have been reported in other studies in Atlantic salmon (Mäkinen et al., 2014; López et al., 2018) and other species (Bahbahani et al., 2015). However, we did find some degree of overlap among genes detected by both haplotype methods and the F_{ST} method as shown in **Figure 7** and **Table 6**.

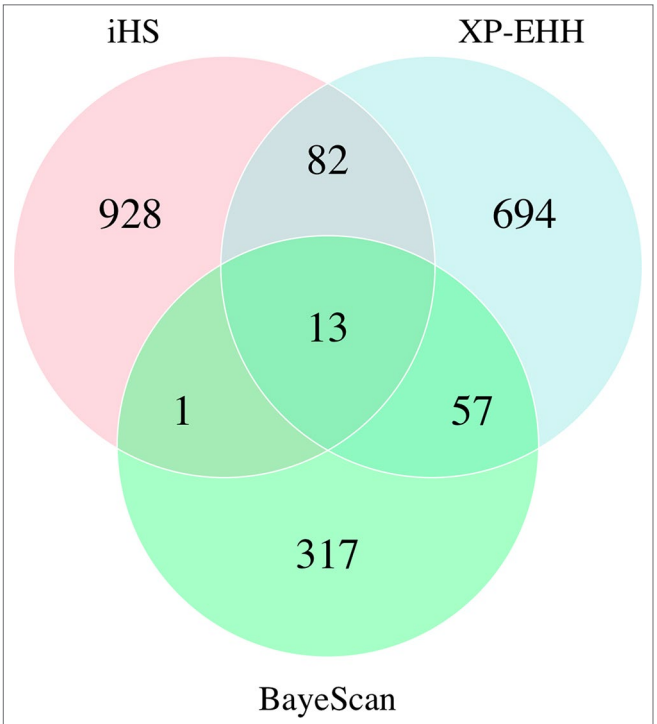


FIGURE 7 | Venn diagram showing shared genes identified among three independent tests in the four populations of Atlantic salmon.

Biological Function of Candidate Selected Regions

Geographical adaptation and selection in farmed Atlantic salmon has resulted in considerable differences between wild and farmed strains (Glover et al., 2009). Genomic regions detected in this study strongly suggest selection on traits that could be associated with either natural or artificial selection, as they relate to the immune system, growth, and behavior, which are all often altered through domestication. Growth has been the main trait focused on by the breeding programs represented by our focal salmon populations. In agreement with this, we found several genes showing evidence of selection that could be potentially influencing growth such as *chp2* and *ccser1*, which were associated with body weight in a previous genome-wide association study (GWAS) on Atlantic salmon (Yoshida et al., 2017). We detected the *kind1* gene that is also associated with growth traits in juvenile, farmed Atlantic salmon (Tsai et al., 2015). It has also been shown that insulin growth factors (IGFs), IGF receptors, and IGF binding proteins, play an important role in regulating growth in several teleost fish species (Duan, 1997). We detected the IGF 1-receptor (*igf1r*), IGF binding protein 6 paralog A2 (*igfbp-6a2*), and IGF binding protein-related protein 1 precursor (*igfbprp1*) as being under selection. We hypothesize that these genes are all contributing to weight variation in farmed salmon. The GO analyses for our

TABLE 6 | Genes detected by at least two selection signatures methods. Genes are indicated in the left column and in the right column their corresponding methods.

GENES	METHODS
CRISP3, NOTCHL, GPSM1B, SI : ZFOS-367G9.1, PHF1, FQ976914.1, TAP1, PBX2, DNASE2, RGL2, PLCL2, SYNGAP1B, BRD2A	iHS; XP-EHH; BayeScan
CRISP3, SI : ZFOS-367G9.1, NOTCHL, GPSM1B, PHF1, FQ976914.1, TAP1, PBX2, DNASE2, RGL2, PLCL2, SYNGAP1B, BRD2A, DOCK10, CRK, LRRC75A, SI : CH211-232I5.3, BLOC1S2, SI : DKEYP-51F12.3, CEP120, CABZ01077978.1, SI : CH211-232I5.1, PRKAA1, PLPP3, BX546500.1, DHCR24, USP24, DAB1A, PRDM5, ANAPC4, SLC10A4, FRYL, PALLD, SLAIN2, MOGAT3B, C1QTNF7, FTR14, LRRC66, SGCB, RASL11B, NDNF, ZBTB34, CPEB2, CC2D2A, FBXL5, NEK1, SH3RF1, OCIAD2, DCUN1D4, USP46, OCIAD1, SCFD2, CDKN1BB, YARS2, PPARAB, BX537249.1, JPH3, KLHDC4, SLC7A5, HMCN2, CDH13, RANBP10, NUTF2, EDC4, NRN1LA, MBTPS1, SLC38A8, PNP6, CALB2A, PSKH1, NECAB2, SCAPER, PSTPIP1A, THBS4A, SERINC5, TRAFD1, SMTNB, UBE2G1B, ANAPC7, ADORA2AA, GUCD1, TAS2R200.1, GSTT1A, DERL3, SMARCB1A, ATP2A2A, BCR, SPECC1LA, SI : CH211-191O15.6, SNRPD3, P2RX7, MMP11A, RALGDS, IFT81, MPEG1.1	iHS; XP-EHH
UNC13B, CRISP3, SI : ZFOS-367G9.1, NOTCHL, GPSM1B, PHF1, FQ976914.1, TAP1, PBX2, DNASE2, RGL2, PLCL2, SYNGAP1B, BRD2A	iHS; BayeScan
NXPH1, ICA1, MIOS, COL28A1B, TAC1, SEPT7B, NEK10, NR1D2B, PHLPP1, RAB5AB, EFHB, CRISP3, SI : ZFOS-367G9.1, NOTCHL, GPSM1B, PHF1, GLCCI1, COL28A1A, RARGA, UBE2E2, ZNF385D, SATB1, FQ976914.1, TAP1, PBX2, DNASE2, RGL2, PLCL2, SYNGAP1B, BRD2A, CTSS2.1, STARD13A, VASH1, OLFM4, RPS6KL1, AREL1, FCF1, ANGEL1, DLST, ESRRB, GPATCH2, TGFB3, PROX2, TMEM179, ARHGEF18B, CABZ01071407.1, ATXN3, SERPINA10, FOXP1A, SI : DKEY-206P8.1, DDX24, SI : CH1073-416D2.4, PRIMA1, UBR7, ITPK1B, HSPA4L, MRPL35, SI : DKEY-21A6.5, CABZ01052815.1, CABZ01066926.1, CHMP3, REEP1, BTBD7, PLK4, MYO1CB, AGBL4, MYL2B, PPP1CC, MTMR3, CUX2B	XP-EHH; BayeScan

candidate genes also showed enrichment for categories related to metabolic and developmental processes, which could certainly affect growth.

Genes functioning in host–pathogen interactions may be targets of natural selection more often than genes from other functional categories (Schlenke and Begun, 2003). The populations used in this study have not been artificially selected for disease resistance; however, we suspect that the culture environment has imposed natural selection on regions implicated in immune system function. We found evidence of selection in seven genes (*kcnb2*, *rlf*, *synrg*, *snx14*, *fbxl5*, *e2f4*, *blm*) that were previously shown to be affected by parasite-driven selection (Zueva et al., 2014). We also identified three genes potentially under selection *(kcnq1*, *lrp5,* and *sh3rf1)* that have were associated with disease resistance in the face of a bacterial disease (*Piscirickettsia salmonis*) in Coho salmon (Barría et al., 2018) and *mettl12* which is associated with immune response to parasites in three-spined stickleback (Huang et al., 2016).

Behavioral traits are among the first traits affected by animal domestication (Kohane and Parsons, 1988), and it has been suggested that domestication may impact behavior even after only one generation (Huntingford, 2004). Among our candidate genes putatively under selection, we identified the endoplasmic reticulum protein 27 (*erp27*) gene, the differential expression of which has been associated to tameness in the red junglefowl (Bélteky et al., 2016). Also, among our candidates were genes, such as *gabrb1*, *scaper*, *clstn3,* and *pex5*, related to mental disorders in humans such alcoholism and schizophrenia (Glatt et al., 2005; Enoch, 2008; Pettem et al., 2013). We think that these genes may be influencing behavior in the salmon populations we studied, and that the artificial selection and domestication could be acting inadvertently on the traits affected by these genes like those that occur in other domestic animals (Clutton-Brock, 1999).

In salmon culture, early sexual maturation has undesired consequences, such as decreased growth and feed conversion efficiency (Good and Davidson, 2016). To avoid these negative effects, maturation is commonly delayed by exposing fish to continuous light, which affects the perception of seasonality and circannual rhythms (Taranger et al., 2010). We would expect then to find genes underlying traits related to maturation rate as showing signs of selection, which we apparently do. One putatively selected gene that we found that may affect maturation rate is *akap13*, which has been shown to play a role in ovarian development in human (Wu et al., 2015), as well as a gene in the AKAP (*akap11*) family, which was previously associated with age to maturity in Atlantic salmon (Barson et al., 2015).

Other interesting genes spanned by regions showing evidence for selection in this study are *hao1*, which is associated with chicken sexual ornaments (comb size), *myo3a*, which is involved in allowing dogs to sense local environmental stimuli (Wang et al., 2013), and *pgbd4,* which is considered a candidate gene involved in adaptation at the regional scale in Atlantic salmon (Bourret et al., 2013) and so could be functioning in adaptation to the aquaculture environment.

CONCLUSIONS

To summarize, in this study we used three different but complementary statistical approaches, iHS, XP-EHH, and BayeScan to detect selection signatures in four farmed Atlantic salmon populations with the same geographical origin, but adapted to different environmental conditions. The methods used in this study were useful for detecting selection signals across populations and allowed us to find genes that could be related to growth, immune system function, and behavior in this species, characters that are commonly influenced by domestication. This study provides potential candidate genes for traits with both biological and economic importance for Atlantic salmon and establishes a strong platform for further studies seeking to better understand how particular genomic variants influence the evolution and cultivation of this species.

ETHICS STATEMENT

The sampling protocol was previously approved by The Comité de Bioética Animal, Facultad de Ciencias Veterinarias y Pecuarias, Universidad de Chile (certificate 29-2014).

AUTHOR CONTRIBUTIONS

ML and JY conceived the research idea. ML drafted the manuscript and carried out the analyses. TL supervised the data analyses and contributed to discussion and writing. TL, AN, JL, RN, and JY reviewed the manuscript. All authors read and approved the final manuscript.

FUNDING

This work has been conceived on the frame of the grant CORFO (11IEI-12843 and 12PIE17669), Government of Chile.

ACKNOWLEDGMENTS

ML acknowledges the National Commission of Scientific and Technologic Research (CONICYT) for the funding through the National PhD funding program. JY is supported by Núcleo Milenio INVASAL funded by Chile's government program, Iniciativa Científica Milenio from Ministerio de Economía, Fomento y Turismo.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00901/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19(9), 1655–1664. doi: 10.1101/gr.094052.109

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andersson, L. (2012). How selective sweeps in domestic animals provide new insight into biological mechanisms. *J. Intern. Med.* 271, 1–14. doi: 10.1111/j.1365-2796.2011.02450.x

Avila, F., Mickelson, J. R., Schaefer, R. J., and McCue, M. E. (2018). Genome-wide signatures of selection reveal genes associated with performance in American quarter horse subpopulations. *Front. Genet.* 9. 9 (249), 1–13 doi: 10.3389/fgene.2018.00249

Bahbahani, H., H. Clifford, D. Wragg, M.N. Mbole-Kariuki, C. Van Tassell, T. Sonstegard 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

Barría, A., Christensen, K. A., Yoshida, G. M., Correa, K., Jedlicki, A., Lhorente, J. P. et al. (2018). Genomic Predictions and Genome-Wide Association Study of Resistance Against Piscirickettsia salmonis in Coho Salmon (Oncorhynchus kisutch); Using ddRAD Sequencing. *G3: Genes, Genomes, Genet.* 8, 1183. doi: 10.1534/g3.118.200053

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P. et al. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature* 528, 405–408. doi: 10.1038/nature16062

Beaumont, M. A., and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* 13, 969–980. doi: 10.1111/j.1365-294X.2004.02125.x

Bélteky, J., Agnvall, B., Johnsson, M., Wright, D., and Jensen, P. (2016). Domestication and tameness: brain gene expression in red junglefowl selected for less fear of humans suggests effects on reproduction and immunology. *R. Soc. Open Sci.* 3, 160033–160033. doi: 10.1098/rsos.160033

Bourret, V., Dionne, M., Kent, M. P., Lien, S., and Bernatchez, L. (2013). Landscape genomics in Atlantic salmon (Salmo salar): searching for gene–environment interactions driving local adaptation. *Evolution* 67, 3469–3487. doi: 10.1111/evo.12139

Browning, B., and Browning, S. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005

Cesconeto, R. J., Joost, S., McManus, M. C., Paiva, S. R., Cobuci, J. A., and Braccini, J. (2017). Landscape genomic approach to detect selection signatures in locally adapted Brazilian swine genetic groups. *Ecol. Evol.* 7, 9544–9556. doi: 10.1002/ece3.3323

Clutton-Brock, J. (1999). *A natural history of domesticated mammals*. Cambridge, UK: Cambridge University Press.

Driscoll, C. A., Macdonald, D. W., and O'Brien, S. J. (2009). From wild animals to domestic pets, an evolutionary view of domestication. *Proc. Natl. Acad. Sci.* 106, 9971–9978. doi: 10.1073/pnas.0901586106

Duan, C. (1997). The insulin-like growth factor system and its biological actions in fish1. *Integr. Comp. Biol.* 37, 491–503. doi: 10.1093/icb/37.6.491

Einum, S., and Fleming, I. (1997). Genetic divergence and interactions in the wild among native, farmed and hybrid Atlantic salmon. *J. Fish Biol.* 50 (3), 634–651. doi: 10.1111/j.1095-8649.1997.tb01955.x

Enoch, M.-A. (2008). The role of GABA(A) receptors in the development of alcoholism. *Pharmacol. Biochem. Behav.* 90, 95–104. doi: 10.1016/j.pbb.2008.03.007

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285–298. doi: 10.1038/hdy.2009.74

FAO (2016). The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all 200 pp.

Foll, M., and Gaggiotti, O. (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

Gautier, M., and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28, 1176–1177. doi: 10.1093/bioinformatics/bts115

Gjedrem, T. (2005). *Selection and Breeding Programs in Aquaculture*. Dordrecht, The Netherlands: Springer. doi: 10.1007/1-4020-3342-7

Gjedrem, T. (2010). The first family-based breeding program in aquaculture. *Rev. Aquacult.* 2, 2–15. doi: 10.1111/j.1753-5131.2010.01011.x

Gjedrem, T. (2012). Genetic improvement for the development of efficient global aquaculture: a personal opinion review. *Aquaculture* 344–349, 12–22. doi: 10.1016/j.aquaculture.2012.03.003

Gjedrem, T., Robinson, N., and Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: a review. *Aquaculture* 350–353, 117–129. doi: 10.1016/j.aquaculture.2012.04.008

Glatt, S. J., Everall, I. P., Kremen, W. S., Corbeil, J., Šášik, R., Khanlou, N., et al. (2005). Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 up-regulation in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15533–15538. doi: 10.1073/pnas.0507666102

Glover, K., H. Ottera, R. Olsen, E. Slinde, G. Taranger and Ø. Skaala (2009). A comparison of farmed, wild and hybrid Atlantic salmon (*Salmo salar* L.) reared under farming conditions. *Aquaculture* 286 (3–4), 203–210. doi: 10.1016/j.aquaculture.2008.09.023

Glover, K. A., Solberg, M. F., McGinnity, P., Hindar, K., Verspoor, E., Coulson, M. W., et al. (2017). Half a century of genetic interaction between farmed and wild Atlantic salmon: status of knowledge and unanswered questions. *Fish Fish. (oxf)* 18, 890–927. doi: 10.1111/faf.12214

Good, C., and Davidson, J. (2016). A review of factors influencing maturation of Atlantic salmon, Salmo salar, with focus on water recirculation aquaculture system environments. *J. World Aquacult. Soc.* 47, 605–632. doi: 10.1111/jwas.12342

Gurgul, A., Jasielczuk, I., Ropka-Molik, K., Semik-Gurgul, E., Pawlina-Tyszko, K., Szmatoła, T., et al. (2018). A genome-wide detection of selection signatures in conserved and commercial pig breeds maintained in Poland. *BMC Genet.* 19, 95–95. doi: 10.1186/s12863-018-0681-0

Gutierrez, A. P., Yáñez, J. M., and Davidson, W. S. (2016). Evidence of recent signatures of selection during domestication in an Atlantic salmon population. *Mar. Genomics* 26, 41–50. doi: 10.1016/j.margen.2015.12.007

Hill, W. G., and Bunger, L. (2004). Inferences on the genetics of quantitative traits from long-term selection in laboratory and domestic animals. *Plant Breed. Rev.* 24, 169–210. doi: 10.1002/9780470650288.ch6

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 2009;4, 44–57. doi: 10.1038/nprot.2008.211

Huang, Y., F. J. J. Chain, M. Panchal, C. Eizaguirre, M. Kalbe, T.L. Lenz et al. (2016). Transcriptome profiling of immune tissues reveals habitat-specific gene expression between lake and river sticklebacks. *Mol. Ecol.* 25, 943–958. doi: 10.1111/mec.13520

Huntingford, F. A. (2004). Implications of domestication and rearing conditions for the behaviour of cultivated fishes. *J. Fish Biol.* 65, 122–142. doi: 10.1111/j.0022-1112.2004.00562.x

Klimentidis, Y. C., Abrams, M., Wang, J., Fernandez, J. R., and Allison, D. B. (2011). Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Hum. Genet.* 129, 407–418. doi: 10.1007/s00439-010-0935-z

Kohane, M., and Parsons, P. (1988). "Domestication," in *Evolutionary biology* Springer, Boston, MA. (Springer), 31–48. doi: 10.1007/978-1-4613-1043-3_2

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533, 200–205. doi: 10.1038/nature17164

Liu, L., Ang, K. P., Elliott, J. A., Kent, M. P., Lien, S., MacDonald, D., et al. (2017). A genome scan for selection signatures comparing farmed Atlantic salmon with two wild populations: testing colocalization among outlier markers, candidate genes, and quantitative trait loci for production traits. *Evol. Appl.* 10, 276–296. doi: 10.1111/eva.12450

López, M. E., Benestan, L., Moore, J. S., Perrier, C., Gilbey, J., Di Genova, A., et al. (2018). Comparing genomic signatures of domestication in two Atlantic salmon (*Salmo salar* L) populations with different geographical origins. *Evol. Appl.* 12(1): 137–156. doi: 10.1111/eva.12689

López, M. E., Neira, R., and Yáñez, J. M. (2015). Applications in the search for genomic selection signatures in fish. *Front. Genet.* 5, 458. doi: 10.3389/fgene.2014.00458

Lorenzen, K., Beveridge, M. C. M., and Mangel, M. (2012). Cultured fish: integrative biology and management of domestication and interactions with wild fish. *Biol. Rev.* 87, 639–660. doi: 10.1111/j.1469-185X.2011.00215.x

Maiorano, A. M., Lourenco, D. L., Tsuruta, S., Ospina, A. M. T., Stafuzza, N. B., Masuda, Y., et al. (2018). Assessing genetic architecture and signatures of selection of dual purpose Gir cattle populations using genomic information. *PLOS One* 13, e0200694. doi: 10.1371/journal.pone.0200694

Mäkinen, H., Vasemägi, A., McGinnity, P., Cross, T. F., and Primmer, C. R. (2014). Population genomic analyses of early-phase Atlantic Salmon (*Salmo salar*) domestication/captive breeding. *Evol. Appl.* 8, 93–107. doi: 10.1111/eva.12230

Manunza, A., Cardoso, T. F., Noce, A., Martínez, A., Pons, A., Bermejo, L. A., et al. (2016). Population structure of eleven Spanish ovine breeds and detection of selective sweeps with BayeScan and hapFLK. *Sci. Rep.* 6, 27296. doi: 10.1038/srep27296

Norris, A. T., Bradley, D. G., and Cunningham, E. P. (1999). Microsatellite genetic variation between and within farmed and wild Atlantic salmon (*Salmo salar*) populations. *Aquaculture* 180 (3–4), 247–264. doi: 10.1016/S0044-8486(99)00212-4

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc B: Biol. Sci.* 365, 185–205. doi: 10.1098/rstb.2009.0219

Pettem, K. L., Yokomaku, D., Luo, L., Linhoff, M. W., Prasad, T., Connor, S. A., et al. (2013). The specific α-neurexin interactor calsyntenin-3 promotes excitatory and inhibitory synapse development. *Neuron* 80, 113–128. doi: 10.1016/j.neuron.2013.07.016

Price, E. O. (1984). Behavioral aspects of animal domestication. *Q. Rev. Biol.* 59, 1–32. doi: 10.1086/413673

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575.

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

Rubin, C-J., H-J. Megens, A. M. Barrio, K. Maqbool, S. Sayyab, D. Schwochow et al. (2012). Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci.* 109 (48), 19529–19536. doi: 10.1073/pnas.1217149109

Ruiz-Larrañaga, O., Langa, J., Rendo, F., Manzano, C., Iriondo, M., and Estonba, A. (2018). Genomic selection signatures in sheep from the Western Pyrenees. *Genet. Sel. Evol.* 50, 9. doi: 10.1186/s12711-018-0378-x

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi: 10.1038/nature01140

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., et al. (2006). Positive natural selection in the human lineage. *Science* 3129, 1614–1620. doi: 10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Schlenke, T. A., and Begun, D. J. (2003). Natural selection drives *Drosophila* immune system evolution. *Genetics* 164, 1471–1480.

Simianer, H., Qanbari, S., and Gianola, D. (2010). "Detection of selection signatures within and between cattle populations," in *Proceedings of 9th World Congress on Genetics Applied to Livestock Production.*

Solberg, M. F., Kvamme, B. O., Nilsen, F., and Glover, K. (2012). Effects of environmental stress on mRNA expression levels of seven genes related to oxidative stress and growth in Atlantic salmon *Salmo salar* L. of farmed, hybrid and wild origin. *BMC Res. Notes* 5 (672), 1–15. doi: 10.1186/1756-0500-5-672

Taranger, G. L., Carrillo, M., Schulz, R. W., Fontaine, P., Zanuy, S., Felip, A., et al. (2010). Control of puberty in farmed fish. *Gen. Comp. Endocrinol.* 165, 483–515. doi: 10.1016/j.ygcen.2009.05.004

Taye, M., Lee, W., Jeon, S., Yoon, J., Dessie, T., Hanotte, O., et al. (2017). Exploring evidence of positive selection signatures in cattle breeds selected for different traits. *Mamm. Genome* 28, 528–541. doi: 10.1007/s00335-017-9715-6

Teletchea, F., and Fontaine, P. (2014). Levels of domestication in fish: implications for the sustainable future of aquaculture. *Fish Fish.* 15, 181–195. doi: 10.1111/faf.12006

Thodesen, J., Grisdale-Helland, B., Helland, S. J., and Gjerde, B. (1999). Feed intake, growth and feed utilization of offspring from wild and selected Atlantic salmon (*Salmo salar*). *Aquaculture* 180, 237–246. doi: 10.1016/S0044-8486(99)00204-5

Tsai, H-Y., Hamilton, A., Tinch, A. E., Guy, D. R., Gharbi, K., Stear, M. J., et al. (2015). Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genomics* 16, 969–969. doi: 10.1186/s12864-015-2117-9

Verspoor, E., Stradmeyer, L., and Nielsen, J. L. (2007). *The Atlantic salmon: genetics, conservation and management*. Blackwell Publishing Ltd, Oxford, UK. John Wiley & Sons. doi: 10.1002/9780470995846

Voight, B., Kudaravalli, S., Wen, X., and Pritchard, J. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72. doi: 10.1371/journal.pbio.0040072

Wang, G. D., Zhai, W., Yang, H. C., Fan, R. X., Cao, X., Zhong, L., et al. (2013). The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat. Commun.* 4, 1860. doi: 10.1038/ncomms2814

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen* 15, 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x

Wu, X., Devine, K., Quaglieri, C., Driggers, P., and Segars, J. (2015). AKAP13 is required for normal murine ovarian development. *Fertil. Steril.* 10, e134. doi: 10.1016/j.fertnstert.2015.07.415

Yáñez, J. M., Naswa, S., López, M. E., Bassini, L., Correa, K., Gilbey, J., et al. (2014). Inbreeding and effective population size in a coho salmon (*Oncorhynchus kisutch*) breeding nucleus in Chile. *Aquaculture.* 420, S15–S19. doi: 10.1016/j.aquaculture.2013.05.028

Yáñez, J. M., S. Naswa, M. E. López, L. Bassini, K. Correa, J. Gilbey et. (2016). Genomewide single nucleotide polymorphism discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. *Mol. Ecol. Resour.* 2016 (4), 1002–1-11. doi: 10.1111/1755-0998.12503

Yoshida, G. M., Lhorente, J. P., Carvalheiro, R., and Yáñez, J. M. (2017). Bayesian genome-wide association analysis for body weight in farmed Atlantic salmon (*Salmo salar* L.). *Anim. Genet.* 48, 698–703. doi: 10.1111/age.12621

Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P., Lien, S., and Primmer, C. R. et al. (2014). Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? *PLoS ONE* 9, e91672. doi: 10.1371/journal.pone.0091672

Conflict of Interest: AN was employed by Marine Harvest, Kindrum, Fanad, C. Donegal, Ireland. JL was employed by company Benchmark Genetics Chile, Puerto Montt, Chile. The remaining 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 October 2019
doi: 10.3389/fgene.2019.00927

OPEN ACCESS

Edited by:
Francesca Bertolini, Technical University of Denmark, Denmark

Reviewed by:
Marco Tolone, University of Palermo, Italy
Emiliano Lasagna, University of Perugia, Italy

***Correspondence:**
David E. MacHugh
david.machugh@ucd.ie

†Present address:
Sam Browett, Ecosystems and Environment Research Centre, School of Environment and Life Sciences, University of Salford, Manchester, United Kingdom
Imtiaz A. S. Randhawa, School of Veterinary Science, University of Queensland, Gatton, QLD, Australia
Dawn J. Howard, Michael P. Mullen, Department of Life and Physical Sciences, Athlone Institute of Technology, Athlone, Co. Westmeath, Ireland
James P. Hanrahan, UCD School of Veterinary Medicine, University College Dublin, Belfield, Dublin, Ireland

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *16 May 2019*
Accepted: *05 September 2019*
Published: *08 October 2019*

Citation:
McHugo GP, Browett S, Randhawa IAS, Howard DJ, Mullen MP, Richardson IW, Park SDE, Magee DA, Scraggs E, Dover MJ, Correia CN, Hanrahan JP and MacHugh DE (2019) A Population Genomics Analysis of the Native Irish Galway Sheep Breed. Front. Genet. 10:927. doi: 10.3389/fgene.2019.00927

A Population Genomics Analysis of the Native Irish Galway Sheep Breed

*Gillian P. McHugo[1], Sam Browett[1†], Imtiaz A. S. Randhawa[2†], Dawn J. Howard[3†], Michael P. Mullen[3†], Ian W. Richardson[4], Stephen D. E. Park[4], David A. Magee[1], Erik Scraggs[1], Michael J. Dover[1], Carolina N. Correia[1], James P. Hanrahan[3†] and David E. MacHugh[1,5]**

[1] Animal Genomics Laboratory, UCD School of Agriculture and Food Science, University College Dublin, Dublin, Ireland, [2] Sydney School of Veterinary Science, University of Sydney, Camden, NSW, Australia, [3] Animal and Grassland Research and Innovation Centre, Athenry, Ireland, [4] IdentiGEN Ltd., Blackrock Business Park, Dublin, Ireland, [5] UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland

The Galway sheep population is the only native Irish sheep breed and this livestock genetic resource is currently categorised as 'at-risk'. In the present study, comparative population genomics analyses of Galway sheep and other sheep populations of European origin were used to investigate the microevolution and recent genetic history of the breed. These analyses support the hypothesis that British Leicester sheep were used in the formation of the Galway. When compared to conventional and endangered breeds, the Galway breed was intermediate in effective population size, genomic inbreeding and runs of homozygosity. This indicates that, although the Galway breed is declining, it is still relatively genetically diverse and that conservation and management plans informed by genomic information may aid its recovery. The Galway breed also exhibited distinct genomic signatures of artificial or natural selection when compared to other breeds, which highlighted candidate genes that may be involved in production and health traits.

Keywords: at-risk breed, conservation genomics, genetic diversity, inbreeding, livestock, selection signature, single nucleotide polymorphism

INTRODUCTION

Sheep were domesticated more than 10,000 years ago and have since been bred for a variety of uses including meat, milk and wool production (Taberlet et al., 2011; Larson and Fuller, 2014; MacHugh et al., 2017). During the last 50 years, the focus of the global sheep industry on only a subset of the 1,400 recorded sheep breeds with enhanced productivity and high-quality outputs has resulted in many locally adapted (local) breeds becoming endangered or extinct (Taberlet et al., 2008; Kijas et al., 2009; Kijas et al., 2012). These breeds are generally considered independent genetic units because crosses are usually not used for further reproduction (Taberlet et al., 2008). Local or heritage livestock breeds are important because they constitute reservoirs of biological diversity different to the major production breeds and that may be important genetic resources for domestic animal species in the face of climate change and increased food requirements in the future (Taberlet et al., 2008; Bowles, 2015). To address these future challenges, it will be possible to use targeted genome editing technologies in livestock. Consequently, functionally important natural sequence variants (NSVs) identified in the genomes of locally adapted native and heritage breeds may become increasingly important for genetic improvement programmes (Wells, 2013; Petersen, 2017; Van Eenennaam, 2017).

The local sheep breeds on the periphery of Northern Europe are recognised as heritage livestock populations that should be conserved and represent important sources of novel genetic diversity accumulated over centuries of microevolution and adaptation to marginal agroecological environments (Tapio et al., 2005). In this regard, the Galway sheep breed is the only surviving sheep breed native to Ireland (Curran, 2010); it was once the principal lowland sheep breed in Ireland but is now considered at-risk by the Food and Agriculture Organization (Food and Agriculture Organization, 2019). The Galway breed therefore represents a useful reservoir of genetic variation for domestic sheep and should be conserved.

The Galway breed is thought to have originated as a composite of indigenous and imported sheep populations, present in Ireland in the mid-19th century, through the breeding endeavours at that time, which were concerned mainly with improved wool production (Hanrahan, 1999). Sheep breeds in Ireland during this period include the important Dishley or New Leicester foundational breed developed by Robert Bakewell (Wykes, 2004). However, it was not until 1923 that a formal Galway herd book was established (Curran, 2010; Food and Agriculture Organization, 2019). Therefore, the range of sheep populations ancestral to the Galway breed in the 18th and 19th centuries, coupled with the possibility of more recent gene flow, poses questions concerning the genetic distinctiveness and admixture history of the breed. In addition, the Galway breed has declined from a peak population size in the 1960s when it was the focus of lowland sheep farming in Ireland (Martin, 1975a; Raftice, 2001; Curran, 2010). By 1994, as defined by the UK Rare Breeds Survival Trust, the Galway breed had reached 'critical' status for sheep breeds with only 300 pedigree breeding ewes registered (Curran, 2010). Since being classed as endangered by the Irish Government in 1998, the number of pedigree Galway sheep has increased due to conservation efforts; however, the breed population size is currently decreasing, raising concerns regarding remaining genetic diversity and the overall viability of the population (Curran, 2010; Food and Agriculture Organization, 2019).

As a local breed with a low census population size, the main threat to the long-term survival of the Galway breed is replacement by more productive commercial breeds, which would further reduce the population size, reduce genetic diversity and increase inbreeding. Other challenges faced by threatened local livestock breeds include poor animal husbandry and management, deliberate or inadvertent crossbreeding and geographical isolation, which increases the risk of extinction (Taberlet et al., 2008; Allendorf et al., 2013). In recent years, with the availability of increasingly powerful genomics technologies, a conservation programme for Galway sheep has been proposed that would leverage molecular genetic information (McHugh et al., 2014). McHugh and colleagues also propose that genome-enabled breeding (genomic selection) could be used in threatened livestock populations to improve production, health and reproduction traits, thereby decelerating replacement by modern breeds (Biscarini et al., 2015). Another strategy could leverage multi-breed or across-breed genomic prediction (Iheshiulor et al., 2016). This approach can increase the accuracy of genomic estimated breeding values for small populations such as the Galway breed, since accurate genomic selection requires large numbers of phenotyped and genotyped animals (Iheshiulor et al., 2016).

To provide information that may be relevant to genetic conservation of the Galway sheep breed, we performed high-resolution population genomics analyses in conjunction with 21 comparator breeds of European origin. These analyses included multivariate analyses of genomic diversity, phylogenetic network graph reconstruction, evaluation of genetic structure and inbreeding, modelling of historical effective population sizes and functional analyses of artificial and natural selection across the Galway sheep genome.

MATERIALS AND METHODS

Galway and Irish Suffolk Sheep DNA Sampling

The Galway and Irish Suffolk sheep DNA samples used for the current survey were generated from peripheral blood samples collected in standard heparinised Vacutainer blood collection tubes (Becton-Dickinson Ltd., Dublin, Ireland). High-quality genomic DNA was then purified from 200 µl of blood from each animal using standard laboratory methods (Howard, 2008). The 49 Galway sheep were sampled from 14 different flocks and pedigree information was consulted to minimise relationship among the animals sampled. The sample size breakdown across the 14 flocks in order of decreasing size is: 6, 6, 5, 5, 5, 5, 3, 3, 3, 3, 2, 1, 1, 1. The flocks were geographically dispersed across County Galway in western Ireland (Howard, 2008). The 55 Irish Suffolk sheep were sampled in approximately equal numbers from two experimental flocks maintained by University College Dublin and Teagasc, the Agriculture and Food Development Authority of Ireland (Howard, 2008).

Additional SNP Data Sources and Data Filtering

Medium-density SNP data were obtained from the International Sheep Genomics Consortium Sheep HapMap Project and consisted of 2,819 sheep from 74 breeds genotyped for 49,034 evenly spaced SNPs using the Illumina® OvineSNP50 BeadChip (Kijas et al., 2012). To focus on the Galway breed, a core sample set of 11 breeds, including the Galway breed, was selected for the primary population genomic analyses (n = 615 animals). This included populations previously examined and known to be more closely related due to their shared European origins (Howard, 2008; Kijas et al., 2012). These comparator populations also included widely used breeds, such as the Merino (MER), and at-risk heritage breeds, such as the Dorset Horn (DSH), Soay (SOA) and Wiltshire (WIL) (Food and Agriculture Organization, 2019). **Figure 1** and **Supplementary Table 1** provide further information on the geographical origins of the 11 breeds used for the core sample set analyses. In addition, **Supplementary Table 1** provides information on an expanded sample set of 22 European and Asian breeds, including the core sample set, used for the phylogenetic tree and network graph reconstructions (n = 1,003).

FIGURE 1 | Map showing the geographical locations where breeds historically originated, adapted from Kijas et al. (2012). The number in brackets indicates the sample size. The breeds shown are the Australian Merino (MER), Border Leicester (BRL), Dorset Horn (DSH), Finnish Landrace (FIN), Galway (GAL), Irish Suffolk (ISF), New Zealand Romney (ROM), Scottish Blackface (SBF), Soay (SOA), Scottish Texel (STX), and Wiltshire (WIL).

The initial data set had already been filtered to remove SNPs with <0.99 call rate, assay abnormality, MAF <0.01, discordant genotypes and inheritance problems (Kijas et al., 2012). The core and extended sample genome-wide SNPs data sets for this study were filtered using PLINK v1.07 (Purcell et al., 2007) to remove SNPs lacking positional information, SNPs unassigned to any chromosome, or SNPs assigned to the X and Y chromosomes (Patterson et al., 2006; Purfield et al., 2012). The final filtered data set was composed of 47,412 SNPs with a total genotyping rate of 99.7%.

Principal Component Analysis

Principal component analysis (PCA) was performed using 47,412 genome-wide SNPs and SMARTPCA from the EIGENSOFT software package (version 4.2) (Patterson et al., 2006). The number of autosomes was set to 26 and breed names were included. The number of outlier removal iterations was set to 0 since outliers could flag individual animals that were the result of crossbreeding. PCA plot visualisations were generated using ggplot2 (Wickham, 2016).

F_{ST} Analysis

Pairwise F_{ST} values (Weir and Cockerham, 1984) were calculated for each pair of breeds using 47,412 genome-wide SNPs and PLINK v1.9 (Chang et al., 2015). Weighted values were chosen to account for different sample sizes for each breed (Weir and Cockerham, 1984).

Construction of Phylogenetic Trees and Ancestry Graphs

Maximum likelihood (ML) phylogenetic trees with ancestry graphs were generated for the core and extended sample data sets using 47,412 genome-wide SNPs and the TreeMix (version 1.12) software package. For the core sample set, the Italian Comisana breed (COM) (Ciani et al., 2014) was used as an outgroup and five migration edges were used for TreeMix visualisation (Pickrell and Pritchard, 2012). The analysis was repeated using the extended sample set of 21 European breeds (**Supplementary Table 1**) and the Indian Garole breed (GAR) was used as an outgroup, again with five migration edges for TreeMix visualisation.

Genetic Structure and Admixture History

Genetic structure and admixture history was investigated for the core sample set of the Galway and 10 other breeds using 47,412 genome-wide SNPs and fastSTRUCTURE (version 1.0) (Raj et al., 2014) as described previously by us (Browett et al., 2018). The analysis was performed with the model complexity, or number of assumed populations, K = 2 to 11. The simple prior approach described by Raj et al. (2014) was used, which is sufficient for modelling population/breed divergence. The 'true' K-value for the number of ancestral populations was estimated using a series of fastSTRUCTURE runs with pre-defined K-values that were examined using the *chooseK.py* script (Raj et al., 2014). Outputs from the fastSTRUCTURE analyses were visualised using the DISTRUCT software program (version 1.1) with standard parameters (Rosenberg, 2004).

Modelling of Current and Historical Effective Population Size

Current and historical effective population size (N_e) trends were modelled with genome-wide SNP linkage disequilibrium data from 47,412 genome-wide SNPs for the core sample set using the *SNeP* software tool (version 1.1) (Barbato et al., 2015) implementing the method for unphased SNP data as described previously by us (Browett et al., 2018). Graphs used to visualise trends in N_e were generated using ggplot2 (Wickham, 2016).

Analysis of Genomic Inbreeding and Runs of Homozygosity

Analysis of genomic inbreeding based on the inbreeding coefficient (F) estimated from SNP heterozygosity data was performed using 47,412 genome-wide SNPs and the PLINK v1.07 –het command (Purcell et al., 2007) since comparable inbreeding results have been observed using pruned or unpruned data for a SNP data set of similar size (Binns et al., 2012).

Runs of homozygosity (ROH) are continuous tracts of homozygosity that most likely arise due to inbreeding and can be identified through surveys of genome-wide SNP data in populations (Curik et al., 2014; Peripolli et al., 2017). Individual animal genomic inbreeding was evaluated as genome-wide autozygosity estimated from the SNP data using runs of homozygosity (ROH) values generated with PLINK v1.07 (Purcell et al., 2007) and the F_{ROH} statistic introduced by McQuillan et al. (2008) with methodologies previously described in detail by Purfield et al. (2012) and Browett et al. (2018). The F_{ROH} statistic represents the proportion of each individual animal's genome covered by ROH, which is generally a consequence of historical inbreeding. Statistical analysis was carried out in R and graphs used to visualise F, F_{ROH} and ROH distributions were generated using ggplot2 (Wickham, 2016; R Core Team, 2018).

Genome-Wide Detection of Signatures of Selection and Functional Enrichment Analysis

The composite selection signal (CSS) method (Randhawa et al., 2014) was used to detect genomic signatures of selection as previously described (Browett et al., 2018). The CSS approach combines the fixation index (F_{ST}), the directional change in selected allele frequency (*ΔSAF*) and cross-population extended haplotype homozygosity (*XP-EHH*) tests into one composite statistic for each SNP in a population genomics data set (Randhawa et al., 2014). For the present study, we used 47,412 genome-wide SNPs genotyped in 49 Galway sheep (GAL) and a sample of 50 randomly selected sheep (5 selected at random from each of the other 10 breeds in the core data set). To mitigate against false positives, genomic selection signatures were only considered significant if at least one SNP from the set of the top 0.1% genome-wide CSS scores was flanked by at least five SNPs from the set of the top 1% CSS scores.

As described previously (Browett et al., 2018), the Ensembl BioMart data mining resource (Smedley et al., 2015) was used to identify genes within ±1.0 Mb of each selection peak (Ensembl release 85, July 2016). Ingenuity® Pathway Analysis (IPA®: Qiagen, Redwood City, CA, USA; release date July 2016) was then used to perform an overrepresentation enrichment analysis with this gene set to identify canonical pathways and functional processes of biological importance. The total gene content of Ensembl release 85 version of the OAR3.1 ovine genome assembly (Jiang et al., 2014) was used as the most appropriate reference gene set for these analyses (Timmons et al., 2015).

RESULTS AND DISCUSSION

Analyses of Breed Divergence, Genetic Differentiation and Admixture

The results of multiple population genomics analyses support the genetic distinctiveness of the Galway sheep population as a discrete breed. The PCA results plotted in **Figure 2** demonstrate separation of the majority of breeds into distinct population clusters, with the notable exceptions of the Australian Merino (MER) and Scottish Blackface (SBF). However, it is important to note that the PCA plot visualisation shown in **Figure 2** did not include the 110 samples from the Soay breed (SOA). A long history as a relatively small isolated island population (Berenos et al., 2016) has led to a marked pattern of genetic differentiation from other breeds, which is evident in the first principal component (PC1) of **Supplementary Figure 1**. Consequently, when the Soay breed is included in a PCA, PC3 is required to separate the Galway breed from the other populations (**Supplementary Figure 2**). Otherwise, the Galway breed clusters with the Scottish Texel breed (STX) and is located close to the Border Leicester breed (BLR). This result supports the documented role for the foundational New Leicester breed in the formation of the Galway and Texel breeds (Porter et al., 2016) and is compatible with the results of a previous study using autosomal microsatellites (Howard, 2008).

The PCA plot shown in **Figure 2** also demonstrated that a number of individual sheep do not cluster closely with other animals from their breeds. This is likely due to recent unacknowledged or inadvertent crossbreeding between animals from different populations (Patterson et al., 2006) or, alternatively, potential mislabelling of particular samples. For example, the 2D

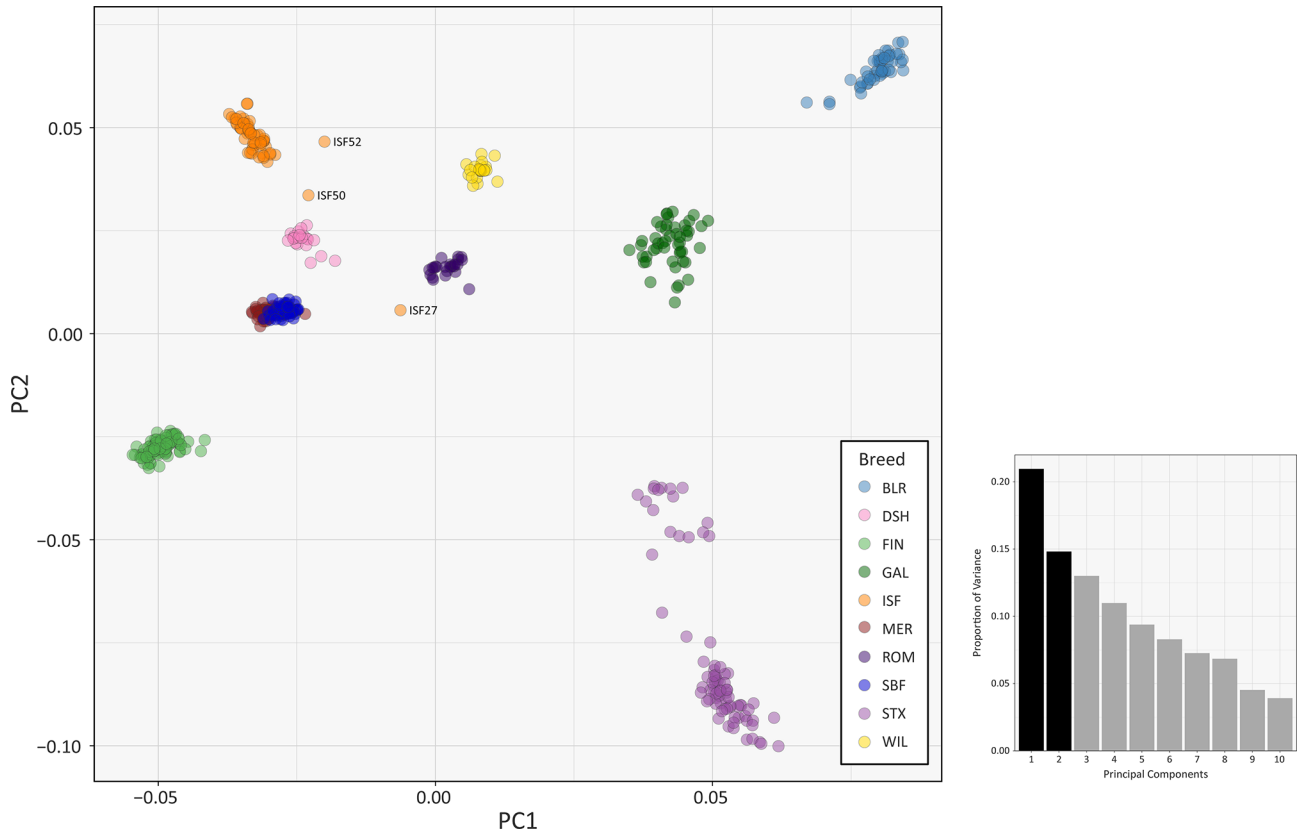


FIGURE 2 | PCA plot generated using 47,412 genome-wide SNPs without the Soay sheep breed (SOA). The first principal component (PC1) is shown on the *x*-axis and the second principal component (PC2) is shown on the *y*-axis. Each breed is designated a different colour and certain individual animals that do not group by breed are labelled. The bar chart shows the proportion of variation explained by each principal component. (For comparison, PC1 versus PC3 is shown in **Supplementary Figure S3** and PC1 versus PC4 is shown in **Supplementary Figure S4**).

and 3D PCA plots shown in **Supplementary Figures 1** and **2** indicate that one of the Irish Suffolk animals (ISF25) was most likely a mislabelled Scottish Texel sample as it emerged within the main Texel cluster for PC1, PC2 and PC3. Consequently, this sample ISF25 was removed from all subsequent analyses.

The PCA results are supported by the interpopulation weighted F_{ST} values for each pair of breeds shown in **Supplementary Table 2**. The results range from 0.080 (Australian Merino and Scottish Blackface) to 0.326 (Soay and Wiltshire). The pairwise F_{ST} values observed for the Galway population sample indicate that, with the exception of the genetically distinctive Soay sheep population (SOA), which inhabits a small island, the breed exhibits moderate genetic differentiation from other European breeds. The Galway breed exhibited relatively low pairwise F_{ST} values with the New Zealand Romney (ROM: 0.110), Australian Merino (MER: 0.118) and Scottish Texel (STX: 0.119) breeds. This is unsurprising because the Romney, Merino and Texel breeds are known to have shared origins with the Galway breed (Curran, 2010; Porter et al., 2016; Food and Agriculture Organization, 2019).

The ML phylogeny and ancestry graph in **Figure 3** shows that the Galway breed groups closely with sheep populations of English and Dutch origin, particularly the Border Leicester (BRL) and the Scottish Texel (STX) breeds. This observation is concordant with previous population genomics studies (Kijas et al., 2012; Fariello et al., 2013) and known breed histories due to the shared historical input of the foundational New Leicester breed (Curran, 2010). The ML phylogeny and ancestry graph generated with additional European breeds and shown in **Supplementary Figure 5** also supports the close relationship among the Galway, BRL and STX breeds. The arrows (graph edges) on **Figure 3** indicate gene flow modelled between populations with the colour scale representing the weight of each migration event.

Results of the genetic structure analysis for individual animals grouped by population are shown in **Figure 4**. Model complexity or numbers of assumed populations (K) ranging from 2 to 11 are visualised to explain the structure in the data and to maximise the marginal likelihood. These results demonstrate that the 11 breeds can be considered discrete populations, thereby supporting interpretation of sheep breeds as separate genetic units (Taberlet et al., 2008) and the genetic distinctiveness of Galway sheep.

The colours on **Figure 4** indicate assignment of individual animals into modelled populations. As with the PCA shown in **Supplementary Figure 1**, the first split (K = 2) separates the isolated Soay sheep population (SOA) from the other breeds. The second split (K = 3) then differentiates the Finnish Landrace (FIN) from the remaining breeds. At K = 9 the Galway breed emerges as

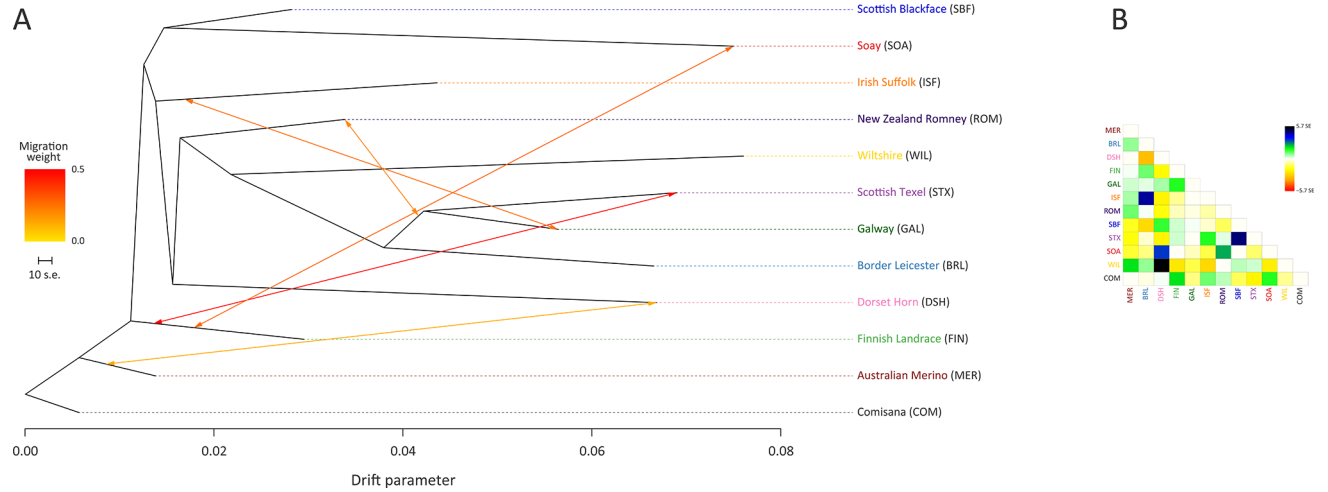


FIGURE 3 | Maximum likelihood (ML) phylogenetic tree network graph generated using 47,412 genome-wide SNPs with five migration edges showing the relationships among 12 sheep breeds **(A)** and the residuals **(B)**. The arrows indicate gene flow events between the populations and the colours of the arrows indicate the relative weights of migration.

a distinct cluster and this genetic component is also apparent in the New Zealand Romney breed (ROM). With $K = 11$ each breed emerges as a distinct genetic cluster. However, some individual animals show evidence of prior crossbreeding or historical admixture, which is indicated by bars that exhibit varying colour proportions. Based on these results, some individual Galway animals exhibit 10% or more admixture with other sheep breeds, particularly the Border Leicester (BRL), Scottish Texel (STX) and Scottish Blackface (SBF). The observed signature of a Galway genomic component in the New Zealand Romney breed (ROM) is supported by the relatively low pairwise F_{ST} value for these breeds, the TreeMix results (**Figure 3**) and their known origins (**Supplementary Table 2**) (Porter et al., 2016).

Modelling Historical Effective Population Size

Figure 5 and **Supplementary Table 4** provide the results of modelling historical effective population size (N_e) for the range of conventional and at-risk sheep breeds (GAL, MER, BRL, DSH, FIN, ISF, ROM, SBF, STX and SOA). Inspection of **Figure 5** and **Supplementary Table 4** shows that the modelled historical trends in N_e for the 11 breeds analysed decline towards the present. However, the GAL breed are intermediate between the breeds with large census populations (FIN, ISF, MER, ROM, SBF and STX) and at-risk breeds with relatively small census populations (BRL, DSH, SOA, WIL) breeds. In addition, the most recent modelled N_e value for the GAL breed is 184 animals 13 generations ago, which is comparable to some of the breeds (e.g. ISF and STX with 178 and 150 animals, respectively). These modelled N_e values, which are based on linkage disequilibrium, may be underestimates due to the physical linkage between many SNPs (Hall, 2016).

To examine these historical trends in N_e more systematically, the data for each breed were shown to be not normally distributed using the Kolmogorov-Smirnov test (**Supplementary Table 3**). Therefore, the non-parametric general Kruskal-Wallis test followed by pairwise Wilcoxon rank sum tests for all population/breed comparisons with adjustment for multiple statistical tests performed with the Bonferroni correction. This analysis demonstrated that the GAL historical N_e trend is significantly different only from the MER breed ($P_{adj.} = 0.006$; **Supplementary Table 5**). Livestock populations tend to exhibit lower N_e values than comparable wild mammal populations (Waples et al., 2016). Notwithstanding this, from a conservation perspective, it is reassuring that the most recent estimated N_e value of 184 for the GAL is above the critical threshold of 100 animals considered essential for the long-term survival of livestock populations (Meuwissen, 2009). This 'demographic fingerprint' (Barbato et al., 2015) is most likely a consequence of the widespread use of the Galway breed for lowland sheep production in Ireland up until the 1980s (Raftice, 2001; Curran, 2010).

Genomic Inbreeding and Runs of Homozygosity

The recent N_e of each of the sheep breeds modelled in **Figure 5** will have been substantially influenced by their inbreeding histories. In this regard, the genomic inbreeding coefficient (F) values estimated for individual animals across all breeds range up to 0.389 for a single Dorset Horn (DSH) animal (**Figure 6**). The majority of F values for individual animals in each breed were not normally distributed based on Shapiro-Wilk test results (**Supplementary Table 3**); therefore, the median F values were generated and evaluated for each breed (**Supplementary Table 6**). The breeds with the highest median F values were the SOA (0.308) and the WIL (0.299) and the two breeds with the lowest median F values were the MER (0.045) and the SBF (0.060). The other breeds exhibited intermediate median F values: BRL (0.243), DSH (0.169), FIN (0.087), GAL (0.127), ISF (0.185), ROM (0.086) and STX (0.111). These results provide a window on the different population histories for the breeds. For example, Soay sheep

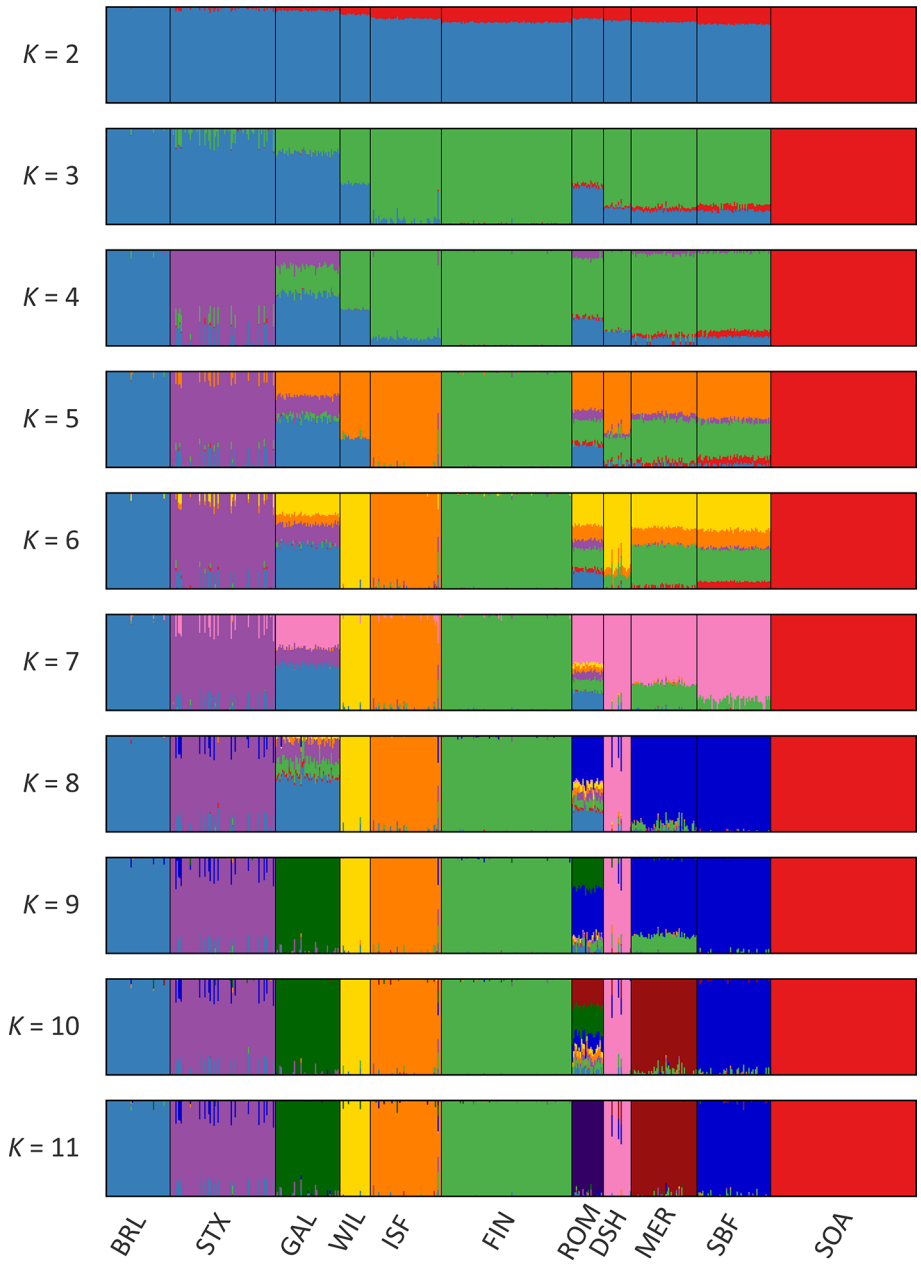


FIGURE 4 | Hierarchical clustering of individual animals using 47,412 genome-wide SNPs. Results are shown for a range of assumed values (K = 2 - 11) for the number of ancestral populations.

(SOA) have existed as a relatively small and isolated population on the island of Soay for hundreds of years while the Wiltshire breed (WIL) has recently experienced a dramatic decline in census population and is considered at risk by the FAO (Food and Agriculture Organization, 2019). From a genetic conservation perspective, except for a single outlier (GAL26), it is encouraging that the Galway breed (GAL) exhibits an intermediate median F value calculated using genome-wide SNP data.

A systematic analysis of F value distributions using the non-parametric Kruskal-Wallis test indicated there were significant differences among breeds (H = 477.33, df = 10, P < 0.001). An analysis of all pairwise breed comparisons using the non-parametric Wilcoxon rank sum test (with Bonferroni correction) was then performed (**Supplementary Table 8**). These results showed that the majority of pairwise comparisons were highly significant, again reflecting the distinct demographic histories of each breed.

Overall, comparable results to those obtained using the genomic inbreeding coefficient (F) were observed for inbreeding coefficients estimated using ROH (F_{ROH}) (**Figure 7**, **Supplementary Tables 3**, **6**, **7** and **9**). However, there were some notable differences; in particular, the lower median F_{ROH} value of 0.101 for the Soay breed (SOA) is likely due to their longer geographical isolation and a

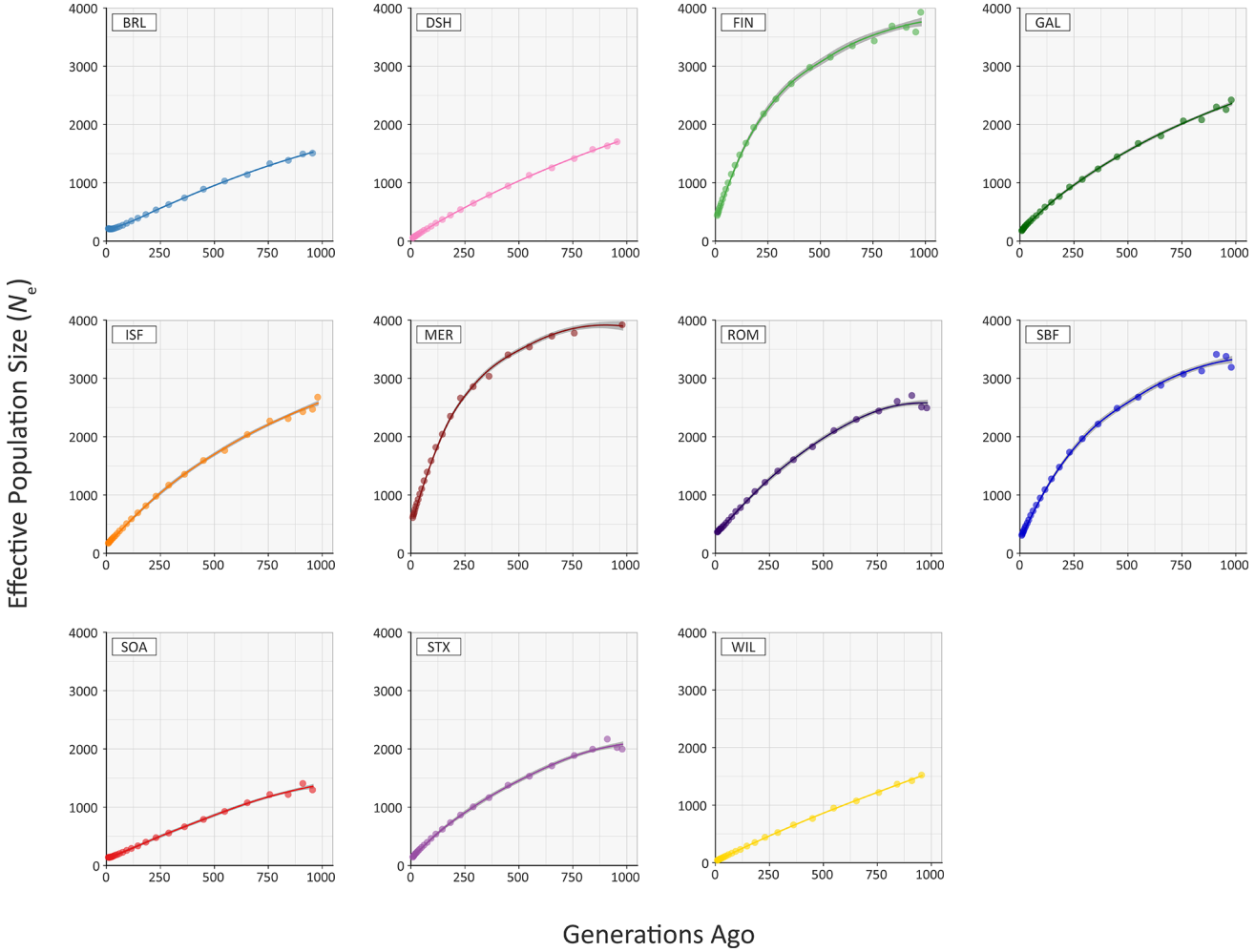


FIGURE 5 | Trends in effective population size (N_e) estimated using 47,412 genome-wide SNPs.

consequence of early historical inbreeding that produced ROH tracts, which have broken down due to recombination (Barrett, 2012; Purfield et al., 2012). It is also notable that the Galway breed contains several individual animals with higher F_{ROH} values (GAL15, GAL16, GAL18, GAL26 and GAL36) indicating that this statistic is useful for identifying animals that should not be prioritised for conservation programmes. With regards to historical inbreeding in the Galway breed (GAL), inbreeding coefficients have previously been calculated using pedigree information for the population in 1969 (F = 0.019; Martin, 1975b), 1999 (F = 0.020; Raftice, 2001) and 2012 (F = 0.023; McHugh et al., 2014). These results indicate that the general trend in inbreeding has been relatively moderate, which may also be reflected in the results obtained using genomic information reported in the present study. It is important to note that monitoring of inbreeding for genetic conservation and management of potentially deleterious recessive genomic variants can be greatly informed through evaluation of ROH parameters using SNP data (Peripolli et al., 2017).

The mean sum of ROH for different length categories varies among the breeds (**Figure 8**); however, none of the breeds exhibit large mean values for the total length of ROH in the 1 to 5 Mb category. This is because the SNP density on the OvineSNP50 BeadChip is too low to accurately detect ROH in this size range and may not accurately estimate F_{ROH} when short segments are included (**Supplementary Table 7**) (Purfield et al., 2012; Ferenčaković et al., 2013). Notwithstanding this limitation, patterns of ROH, which reflect both recent and older inbreeding histories, are evident. For example, the Wiltshire breed (WIL) has large mean total ROH lengths for the other categories, presumably reflecting both historical and recent inbreeding. Other breeds, such as the Australian Merino (MER), have smaller mean total lengths of ROH in all categories, an observation that is concordant with the results of the genomic inbreeding and the analysis of N_e estimates. This is because individual animals from breeds with larger effective population sizes—such as the Australian Merino—are less likely to be the result of inbreeding and are therefore less likely to contain large ROH segments in their genomes (Curik et al., 2014; Peripolli et al., 2017). The converse of this is true for breeds with lower N_e values and large ROH tracts in their genomes, such as the endangered Wiltshire breed. In terms of mean total

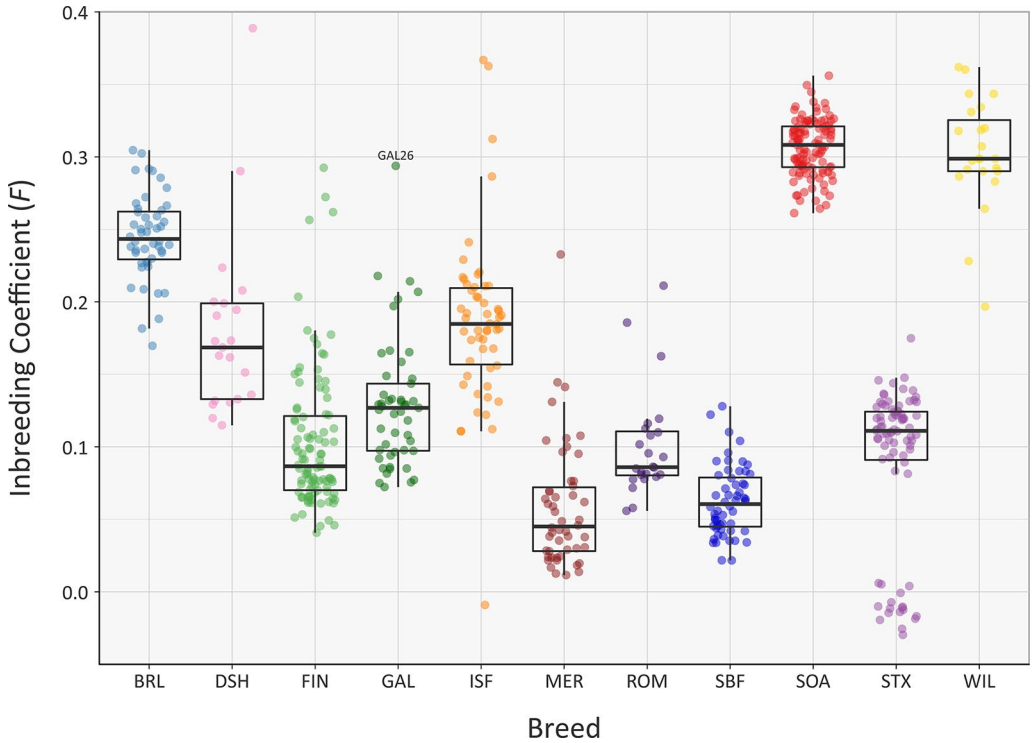


FIGURE 6 | Tukey box plots showing the distribution of *F* values, estimated using 47,412 genome-wide SNPs, for the Galway sheep breed (GAL) and 10 comparator breeds. The single GAL26 outlier is labelled.

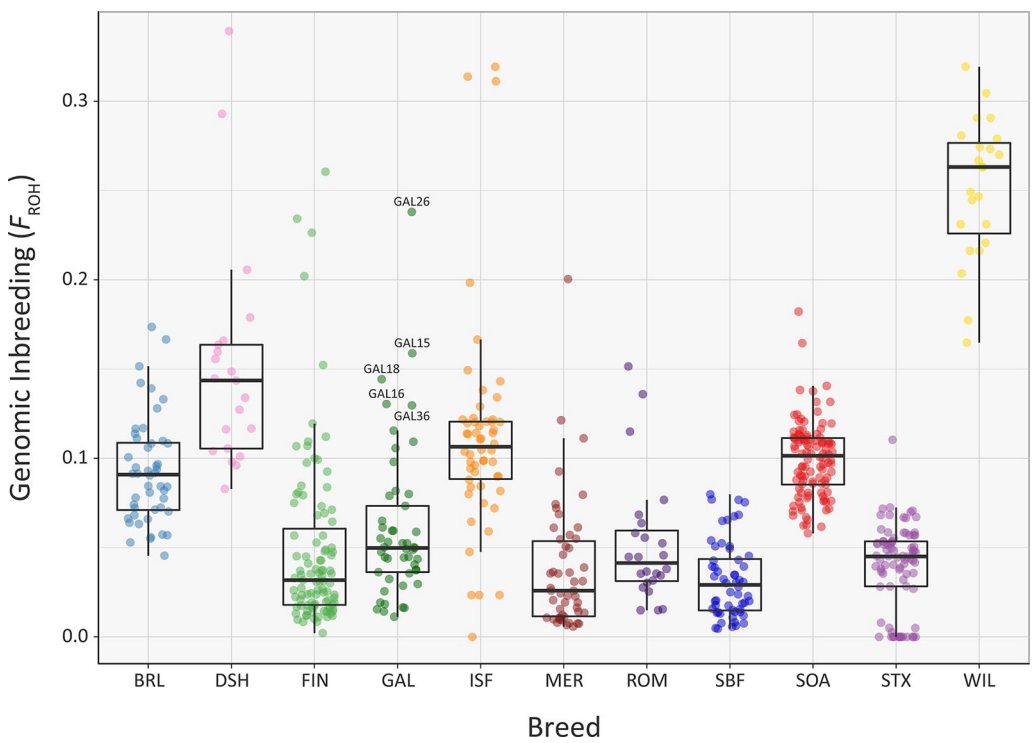


FIGURE 7 | Tukey box plots showing the distribution of F_{ROH} values estimated using 47,412 genome-wide SNPs, for the Galway sheep breed (GAL) and 10 comparator sheep breeds. Five outlier GAL animals are labelled.

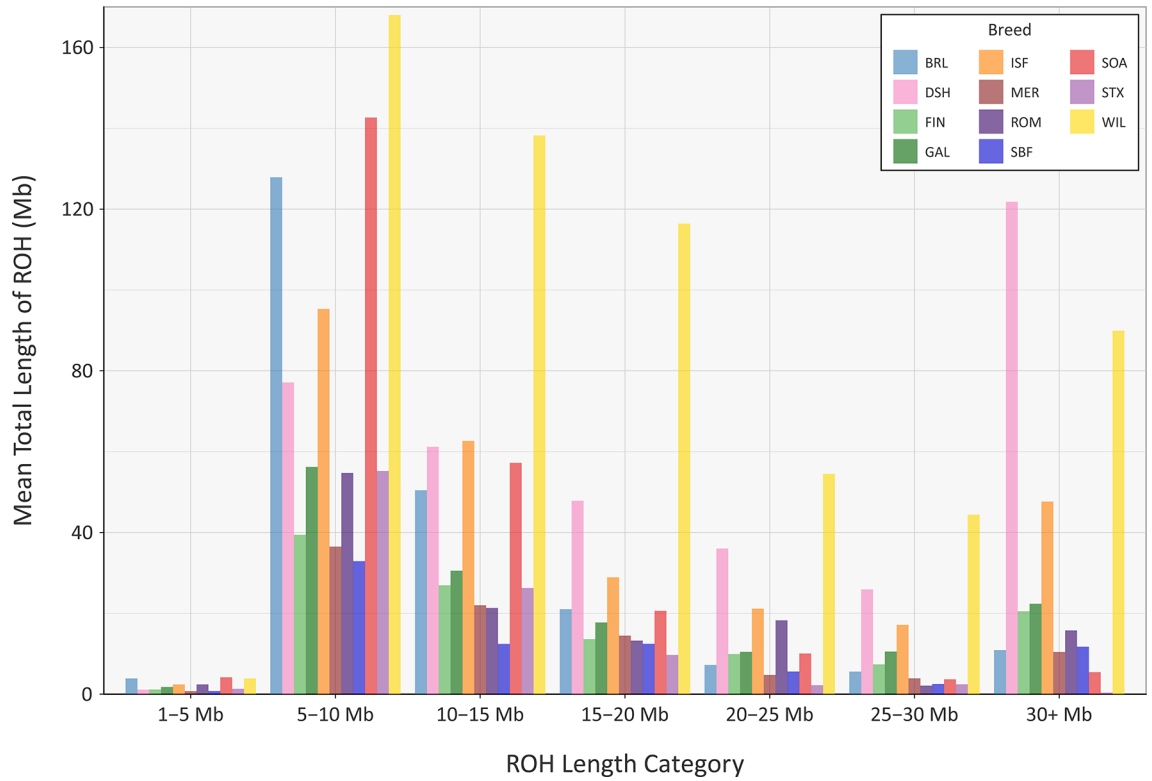


FIGURE 8 | Bar graph showing the mean total length of runs of homozygosity (ROH) in different tract length categories for the Galway sheep breed (GAL) and 10 comparator sheep breeds.

length of ROH, the Galway breed emerges between these extremes, reflecting an intermediate effective population size and history of moderate inbreeding (**Figure 8**). In conjunction with the other analyses of genomic diversity, these results are also encouraging for genetic conservation and the long-term viability of the breed.

Signatures of Selection in the Galway Sheep Breed

Using defined criteria, five significant peaks of selection were detected with the CSS approach (**Figure 9**): two on OAR1, one on OAR3 and two on OAR8 (that merge into one peak on the graph). Each selection peak was located in a ROH tract detected in at least three Galway samples, which may indicate reduced genetic diversity in these regions as a consequence of localised selective sweeps (Purfield et al., 2017). Detection of these selection peaks demonstrates that the Galway population has experienced a unique history of both natural and human-mediated selection, presumably because of adaptation to the agroecology of Ireland, a large Northwestern European island with a temperate oceanic climate.

The precise locations of the peaks that have clusters of SNPs within the top 0.1% CSS score class are provided with additional information in **Supplementary Table 10**. The 197 genes within these regions are listed in **Supplementary Table 11**. Using IPA®, the top five physiological system development and function pathways enriched for the subset of 119 genes that could be mapped to HGNC symbols were identified and are listed in **Table 1** (Krämer et al., 2013).

Of the 119 candidate genes hypothesised to be under selection in the Galway breed, 28 are involved in tissue development and 15 are involved in connective tissue development and function. This is a common observation in studies of selection across the genomes of livestock populations (de Simoni Gouveia et al., 2014; Gutiérrez-Gil et al., 2015; Randhawa et al., 2016). Seven of the 119 genes are involved in hair and skin development and function, which may be explained by the use of Galway sheep in wool production (Curran, 2010). Selection and maintenance of traits that confer resilience to infectious disease is important in domestic animal populations, including many sheep breeds (Bishop and Woolliams, 2014; Bishop, 2015). Thirteen of the 119 genes under the selection peaks are involved in immune cell trafficking, which may be as a result of the climate and unique disease challenges posed by the Irish environment, such as the prevalence of liver fluke (Toolan et al., 2015). A large group of 26 genes enriched for haematological system development and function were also located under the selection peaks; however, a microevolutionary explanation for this is not hypothesised here.

Genetic Conservation of the Galway Sheep Breed

The results of the population genomics analyses presented here are mutually consistent and highlight the utility of

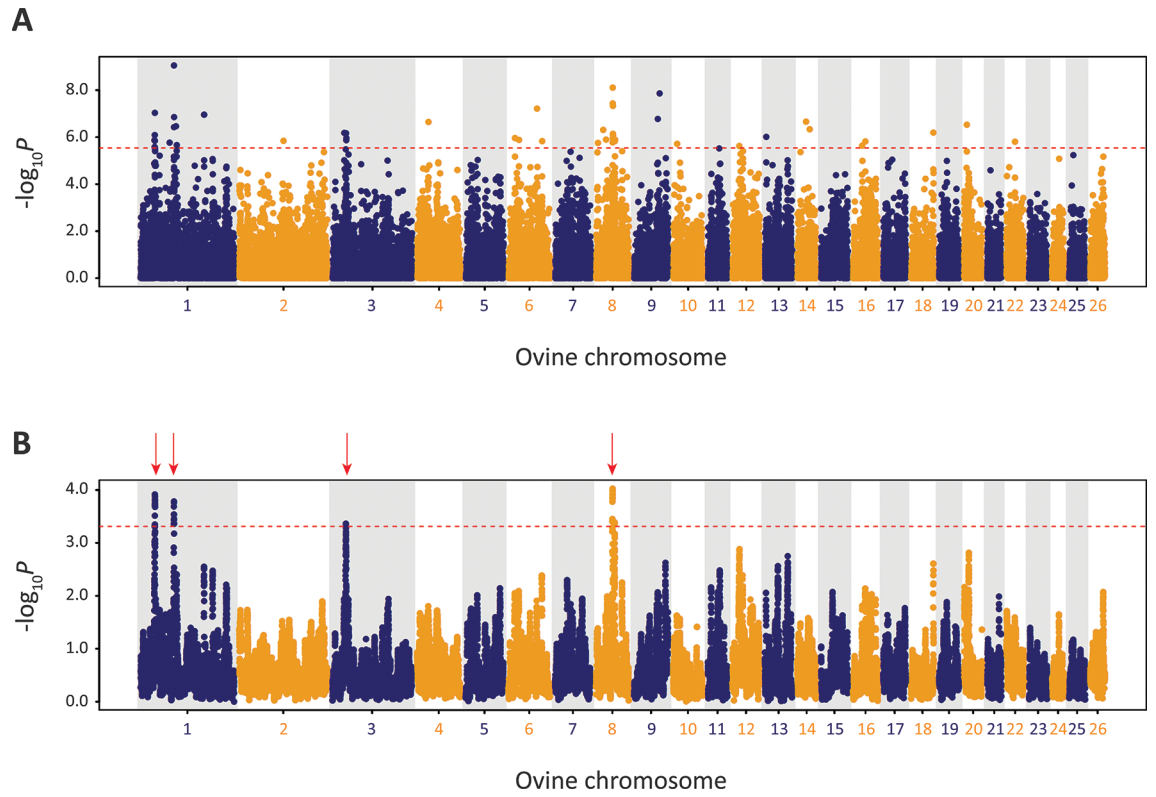


FIGURE 9 | Manhattan plots of composite selection signal (CSS) results for Galway sheep (n = 49) contrasted with a random group selected from the other 10 breeds in the core data set (n = 50). **(A)** Unsmoothed results. **(B)** Smoothed results obtained by averaging CSS of SNPs within each 1Mb window. Red dotted line on each plot denotes the genome-wide 0.1% threshold for the empirical CSS scores. Red vertical arrows indicate selection peaks detected on OAR1, OAR3 and OAR8.

TABLE 1 | Top five physiological system development and function pathways enriched for the 119 candidate genes proximal to the five detected selection peaks.

Pathway	No. of Genes	Range of *P*-values
Tissue development	28	0.037–0.000
Haematological system development and function	26	0.037–0.000
Hair and skin development and function	7	0.016–0.000
Immune cell trafficking	13	0.037–0.001
Connective tissue development and function	15	0.037–0.001

dense genome-wide marker data for conservation genomics in livestock populations, particularly for at-risk heritage landrace populations such as the Galway breed. Our results show the Galway breed is genetically distinct from other European sheep breeds, emerging in multivariate PCA and phylogenetic tree network graph visualisations as a distinct group but close to the Border Leicester breed (BRL), which has been observed previously (Kijas et al., 2012). In terms of effective population size and genomic inbreeding, the Galway breed emerged as intermediate between non-endangered and endangered sheep breeds. This indicates that there is substantial genetic diversity remaining in the population, which could be managed with a conservation programme that is informed by genomic information.

DATA AVAILABILITY STATEMENT

The Galway sheep (GAL) and additional sheep breed Illumina® OvineSNP50 BeadChip data are available as part of the International Sheep Genomics Consortium Ovine SNP50 HapMap Dataset (www.sheephapmap.org/download.php).

ETHICS STATEMENT

Animal biological sample collection was conducted under license issued in accordance with Irish and European Union legislation (Cruelty to Animals Act, 1876, and European Community Directive, 86/609/EC) as described previously (Mullen et al., 2013). All animals were managed in accordance with the guidelines for the accommodation and care of animals under Article 5 of the Directive.

AUTHOR CONTRIBUTIONS

DEM, DH, MM and JH conceived and designed the project. DH, MM, DAM, ES and JH organised sample collection and

genotyping. GM, SB, IAR, IWR, SP, MD and CC performed the analyses. GM and DEM wrote the manuscript and all authors reviewed and approved the final manuscript.

FUNDING

This work was supported by Department of Agriculture, Food and the Marine (DAFM) funding under the Genetic Resources for Food and Agriculture scheme (grant no: 09/GR/06); an Investigator Programme Grant from Science Foundation Ireland (SFI/08/IN.1/B2038); a Research Stimulus Grant from DAFM (RSF 06 406); a European Union Framework 7 Project Grant (KBBE-211602-MACROSYS); the Brazilian Science Without Borders Programme (CAPES grant no. BEX-13070-13-4); and the UCD MSc Programme in Evolutionary Biology.

ACKNOWLEDGMENTS

The authors would like to thank Jon Yearsley, Kay Nolan, John Leech, Lorenzo de Jonge, Charlotte Gilbert and Han Haige for helpful discussion and feedback. We also thank members of the Galway Sheep Breeders Association for their cooperation in sample and data collection. Genotypic data used in this study was collected as part of a coordinated international project run under the auspices of the International Sheep Genomics Consortium.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00927/full#supplementary-material

REFERENCES

Allendorf, F. W., Luikart, G., and Aitken, S. N. (2013). *Conservation and the genetics of populations*. Oxford, UK: John Wiley & Sons.

Barbato, M., Orozco-terWengel, P., Tapio, M., and Bruford, M. W. (2015). *SNeP*: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front. Genet.* 6, 109. doi: 10.3389/fgene.2015.00109

Barrett, R. D. (2012). Bad coat, ripped genes: cryptic selection on coat colour varies with ontogeny in Soay sheep. *Mol. Ecol.* 21, 2833–2835. doi: 10.1111/j.1365-294X.2012.05560.x

Berenos, C., Ellis, P. A., Pilkington, J. G., and Pemberton, J. M. (2016). Genomic analysis reveals depression due to both individual and maternal inbreeding in a free-living mammal population. *Mol. Ecol.* 25, 3152–3168. doi: 10.1111/mec.13681

Binns, M. M., Boehler, D. A., Bailey, E., Lear, T. L., Cardwell, J. M., and Lambert, D. H. (2012). Inbreeding in the Thoroughbred horse. *Anim. Genet.* 43, 340–342. doi: 10.1111/j.1365-2052.2011.02259.x

Biscarini, F., Nicolazzi, E. L., Stella, A., Boettcher, P. J., and Gandini, G. (2015). Challenges and opportunities in genetic improvement of local livestock breeds. *Front. Genet.* 6, 33. doi: 10.3389/fgene.2015.00033

Bishop, S. C. (2015). Genetic resistance to infections in sheep. *Vet. Microbiol.* 181, 2–7. doi: 10.1016/j.vetmic.2015.07.013

Bishop, S. C., and Woolliams, J. A. (2014). Genomics and disease resistance studies in livestock. *Livestock Sci.* 166, 190–198. doi: 10.1016/j.livsci.2014.04.034

Bowles, D. (2015). Recent advances in understanding the genetic resources of sheep breeds locally-adapted to the UK uplands: opportunities they offer for sustainable productivity. *Front. Genet.* 6, 24. doi: 10.3389/fgene.2015.00024

Browett, S., McHugo, G., Richardson, I. W., Magee, D. A., Park, S. D. E., Fahey, A. G., et al. (2018). Genomic characterisation of the indigenous Irish Kerry cattle breed. *Front. Genet.* 9, 51. doi: 10.3389/fgene.2018.00051

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Ciani, E., Crepaldi, P., Nicoloso, L., Lasagna, E., Sarti, F. M., Moioli, B., et al. (2014). Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. *Anim. Genet.* 45, 256–266. doi: 10.1111/age.12106

Curik, I., Ferenčaković, M., and Sölkner, J. (2014). Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livestock Sci.* 166, 26–34. doi: 10.1016/j.livsci.2014.05.034

Curran, P. L. (2010). *The Native Lowland Sheep of Galway and Roscommon: a history*. Tara, Co. Meath, Ireland: Patrick Leonard Curran.

de Simoni Gouveia, J. J., da Silva, M. V., Paiva, S. R., and de Oliveira, S. M. (2014). Identification of selection signatures in livestock species. *Genet. Mol. Biol.* 37, 330–342. doi: 10.1590/S1415-47572014000300004

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

Ferenčaković, M., Sölkner, J., and Curik, I. (2013). Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genet. Sel. Evol.* 45, 42. doi: 10.1186/1297-9686-45-42

Food and Agriculture Organization. (2019). Domestic Animal Diversity Information System (DAD-IS); available at: www.fao.org/dadis. [Accessed 14/02/2019].

Gutiérrez-Gil, B., Arranz, J. J., and Wiener, P. (2015). An interpretive review of selective sweep studies in *Bos taurus* cattle populations: identification of unique and shared selection signals across breeds. *Front. Genet.* 6, 167. doi: 10.3389/fgene.2015.00167

Hall, S. J. (2016). Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. *Animal* 10, 1778–1785. doi: 10.1017/S1751731116000914

Hanrahan, J. P. 1999. The Galway Breed – Origins and the future; available at www.galwaysheep.ie. [Accessed 13/07/2018].

Howard, D. J. (2008). *A comparative study of molecular genetic variation in three Irish sheep breeds; the Galway, Texel and Suffolk*. University College Dublin: PhD thesis.

Iheshiulor, O. O., Woolliams, J. A., Yu, X., Wellmann, R., and Meuwissen, T. H. (2016). Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet. Sel. Evol.* 48, 15. doi: 10.1186/s12711-016-0193-1

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J. F., Faraut, T., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 344, 1168–1173. doi: 10.1126/science.1252806

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, 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

Kijas, J. W., Townley, D., Dalrymple, B. P., Heaton, M. P., Maddox, J. F., McGrath, A., et al. (2009). A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS One* 4, e4668. doi: 10.1371/journal.pone.0004668

Krämer, A., Green, J., Pollard, J. Jr., and Tugendreich, S. (2013). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30, 523–530. doi: 10.1093/bioinformatics/btt703

Larson, G., and Fuller, D. Q. (2014). The evolution of animal domestication. *Annu. Rev. Ecol. Evol. Syst.* 45, 115–136. doi: 10.1146/annurev-ecolsys-110512-135813

MacHugh, D. E., Larson, G., and Orlando, L. (2017). Taming the past: ancient DNA and the study of animal domestication. *Annu. Rev. Anim. Biosci.* 5, 329–351. doi: 10.1146/annurev-animal-022516-022747

Martin, I. (1975a). A genetic analysis of the Galway sheep breed: 1. some aspects of population dynamics of the pedigree and non-pedigree Galway sheep breed. *Ir. J. Agric. Sci.* 14, 245–253.

Martin, I. (1975b). A genetic analysis of the Galway sheep breed: 3. level of inbreeding in the pedigree Galway sheep breed. *Ir. J. Agric. Sci.* 14, 269–274.

McHugh, N., Berry, D., McParland, S., Wall, E., and Pabiou, T. (2014). *Irish sheep breeding: current status and future plans*. (Cork, Ireland: Teagasc and Sheep Ireland).

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., et al. (2008). Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 83, 359–372. doi: 10.1016/j.ajhg.2008.08.007

Meuwissen, T. (2009). Genetic management of small populations: a review. *Acta Agric. Scand. A Anim. Sci.* 59, 71–79. doi: 10.1080/09064700903118148

Mullen, M. P., Hanrahan, J. P., Howard, D. J., and Powell, R. (2013). Investigation of prolific sheep from UK and Ireland for evidence on origin of the mutations in *BMP15* (*Fec*X^G, *Fec*X^B) and *GDF9* (*Fec*G^H) in Belclare and Cambridge sheep. *PLoS One* 8, e53172. doi: 10.1371/journal.pone.0053172

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2, e190. doi: 10.1371/journal.pgen.0020190

Peripolli, E., Munari, D. P., Silva, M., Lima, A. L. F., Irgang, R., and Baldi, F. (2017). Runs of homozygosity: current knowledge and applications in livestock. *Anim. Genet.* 48, 255–271. doi: 10.1111/age.12526

Petersen, B. (2017). Basics of genome editing technology and its application in livestock species. *Reprod. Domest. Anim.* 52 Suppl 3, 4–13. doi: 10.1111/rda.13012

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8, e1002967. doi: 10.1371/journal.pgen.1002967

Porter, V., Alderson, L., Hall, S. J. G., and Sponenberg, D. P. (2016). *Mason's world encyclopedia of livestock breeds and breeding*. Wallingford, UK: CABI. doi: 10.1079/9781845934668.0000

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

Purfield, D. C., Berry, D. P., McParland, S., and Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genet.* 13, 70. doi: 10.1186/1471-2156-13-70

Purfield, D. C., McParland, S., Wall, E., and Berry, D. P. (2017). The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. *PLoS One* 12, e0176780. doi: 10.1371/journal.pone.0176780

R Core Team. 2018. R: a language and environment for statistical computing. Available: https://www.R-project.org.

Raftice, M. J. (2001). *Genetic conservation of the Galway sheep breed*. (Dublin, Ireland: University College Dublin).

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Randhawa, I. A., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet.* 15, 34. doi: 10.1186/1471-2156-15-34

Randhawa, I. A., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2016). A meta-assembly of selection signatures in cattle. *PLoS One* 11, e0153013. doi: 10.1371/journal.pone.0153013

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43, W589–W598. doi: 10.1093/nar/gkv350

Taberlet, P., Coissac, E., Pansu, J., and Pompanon, F. (2011). Conservation genetics of cattle, sheep, and goats. *C. R. Biol.* 334, 247–254. doi: 10.1016/j.crvi.2010.12.007

Taberlet, P., Valentini, A., Rezaei, H., Naderi, S., Pompanon, F., Negrini, R., et al. (2008). Are cattle, sheep, and goats endangered species? *Mol. Ecol.* 17, 275–284. doi: 10.1111/j.1365-294X.2007.03475.x

Tapio, M., Tapio, I., Grislis, Z., Holm, L. E., Jeppsson, S., Kantanen, J., et al. (2005). Native breeds demonstrate high contributions to the molecular variation in northern European sheep. *Mol. Ecol.* 14, 3951–3963. doi: 10.1111/j.1365-294X.2005.02727.x

Timmons, J. A., Szkop, K. J., and Gallagher, I. J. (2015). Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol.* 16, 186. doi: 10.1186/s13059-015-0761-7

Toolan, D. P., Mitchell, G., Searle, K., Sheehan, M., Skuce, P. J., and Zadoks, R. N. (2015). Bovine and ovine rumen fluke in Ireland—Prevalence, risk factors and species identity based on passive veterinary surveillance and abattoir findings. *Vet. Parasitol.* 212, 168–174. doi: 10.1016/j.vetpar.2015.07.040

Van Eenennaam, A. L. (2017). Genetic modification of food animals. *Curr. Opin. Biotechnol.* 44, 27–34. doi: 10.1016/j.copbio.2016.10.007

Waples, R. K., Larson, W. A., and Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity (Edinb)* 117, 233–240. doi: 10.1038/hdy.2016.60

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 36, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x

Wells, K. D. (2013). Natural genotypes via genetic engineering. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16295–16296. doi: 10.1073/pnas.1315623110

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York, USA: Springer Publishing Company.

Wykes, D. L. (2004). Robert Bakewell (1725-1795) of Dishley: farmer and livestock improver. *Agric. Hist. Rev.* 52, 38–55.

Conflict of Interest: The authors IWR and SP are employed by IdentiGEN, Ltd. All other authors declare no competing interests and that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2019
doi: 10.3389/fgene.2019.01025

A Combined Multi-Cohort Approach Reveals Novel and Known Genome-Wide Selection Signatures for Wool Traits in Merino and Merino-Derived Sheep Breeds

Sami Megdiche[1,2], Salvatore Mastrangelo[3], Mohamed Ben Hamouda[4], Johannes A. Lenstra[5] and Elena Ciani[2]**

[1] *Département des Ressources Animales, Agroalimentaire et Développement Rural, Institut Supérieur Agronomique de Chott-Mariem, Université de Sousse, Sousse, Tunisia,* [2] *Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, University of Bari "Aldo Moro," Bari, Italy,* [3] *Dipartimento di Scienze Agrarie, Alimentari e Forestali, University of Palermo, Palermo, Italy,* [4] *INRA-Tunisie, Institute for Risk Assessment Sciences, Ariana, Tunisia,* [5] *Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands*

OPEN ACCESS

Edited by:
Francesca Bertolini, Technical University of Denmark, Denmark

Reviewed by:
Vincenzo Landi, Universidad de Córdoba,Spain
Michelle Mousel, United States Department of Agriculture,United States

***Correspondence:**
Sami Megdiche
megdichesami@hotmail.fr
Elena Ciani,
elena.ciani@uniba.it

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *28 April 2019*
Accepted: *24 September 2019*
Published: *25 October 2019*

Citation:
Megdiche S, Mastrangelo S, Ben Hamouda M, Lenstra JA and Ciani E (2019) A Combined Multi-Cohort Approach Reveals Novel and Known Genome-Wide Selection Signatures for Wool Traits in Merino and Merino-Derived Sheep Breeds. Front. Genet. 10:1025. doi: 10.3389/fgene.2019.01025

Merino sheep represents a valuable genetic resource worldwide. In this study, we investigated selection signatures in Merino (and Merino-derived) sheep breeds using genome-wide SNP data and two different approaches: a classical F_{ST}-outlier method and an approach based on the analysis of local ancestry in admixed populations. In order to capture the most reliable signals, we adopted a combined, multi-cohort approach. In particular, scenarios involving four Merino breeds (Spanish Merino, Australian Merino, Chinese Merino, and Sopravissana) were tested *via* the local ancestry approach, while nine pair-wise breed comparisons contrasting the above breeds, as well as the Gentile di Puglia breed, with non-Merino breeds from the same geographic area were tested *via* the F_{ST}-outlier method. Signals observed using both methods were compared with genome-wide patterns of distribution of runs of homozygosity (ROH) islands. Novel and known selection signatures were detected. The most reliable signals were observed on OAR 3 (*MSRB3* and *LEMD3*), OAR10 (*FRY* and *RXFP2*), OAR 13 (*RALY*), OAR17 (*FAM101A*), and OAR18 (*NFKBIA, SEC23A*, and *PAX9)*. All the above overlapped with known QTLs for wool traits, and evidences from the literature of their involvement in skin/hair/wool biology, as well as gene network analysis, further corroborated these results. The signal on OAR10 also contains well known evidence for association with horn morphology and polledness. More elusive biological evidences of association with the Merino phenotype were observed for a number of other genes, notably *LOC101120019* and *TMEM132B* (OAR17), *LOC105609948* (OAR3), *LOC101110773* (OAR10), and *EIF2S2* (OAR17). Taken together, the above results further contribute to decipher the genetic basis underlying the Merino phenotype.

Keywords: Merino sheep breeds, wool, genome-wide selection signatures, F_{ST}-outlier, local ancestry in admixed populations, runs of homozygosity

INTRODUCTION

Sheep were among the first livestock species to be domesticated (Ryder 1981). Archeological evidences suggest domestication occurred in a region extending from the northern Zagros to southeastern Anatolia ca. 11,000 B.P. (Zeder, 2008). In the last two decades, information from molecular data, as well as discovery and study of novel archaeological sites, has shed new light on the origins and subsequent diffusion of domestic sheep worldwide (Chessa et al., 2009; Meadows et al., 2011; Kijas et al., 2012; Demirci et al., 2013; Singh et al., 2013; Dymova et al., 2017; Ethier et al., 2017; Ivanova et al., 2018a; Ivanova et al., 2018b). Early domesticated sheep are known to have been transported over long distances or even by sea, as early as around 12,000 years BP (Zeder, 2008). They are supposed to have been initially reared mainly for meat and, only during the fifth millennium B.P., specialization for "secondary" products, such as milk and wool, is thought to have occurred (Debono Spiteri et al., 2016). In particular, analysis of viral retro-types combined with archaeological evidence provide support to the hypothesis that specialized wool sheep populations were developed in South-West Asia and then spread throughout Europe, replacing, in the majority of areas, the more primitive domestic stocks (Chessa et al., 2009). Specialization for wool production culminated, in the Middle Ages, with the development of the Merino sheep in Spain. In a recent paper, by analyzing genome-wide SNP data from an intercontinental set of sheep breeds, inclusive of 12 Merino and Merino-derived populations, our group contributed to the reconstruction of the history of Merino development, and the subsequent worldwide merinization process (Ciani et al., 2015).

Well renowned for its premium white fleece and the abundant production of soft, fine, and curly wool, Merino sheep represent a valuable genetic resource worldwide. As such, deciphering the genetic basis underlying the peculiar Merino phenotype is a fundamental aim, and it may further contribute improving wool performances of Merino and Merino-derived breeds. A number of papers have addressed this issue, looking at the genome in search for QTLs (quantitative trait loci) related to wool traits, by using STR markers in Merino (Beh et al., 2001; Bidinost et al., 2008; Roldan et al., 2010), Merino crosses (Rogers et al., 1994; Henry et al., 1998; Zhai et al., 2019), and non-Merino (Allain et al., 1998; Ponz et al., 2001; Allain et al., 2006) sheep populations, or looking at candidate genes (Ling et al., 2014; Rong et al., 2015; Ma et al., 2017; Mu et al., 2017), with keratin genes being among the most studied targets (Parsons et al., 1994; Phuaa et al., 2015; Chai et al., 2017; Li et al., 2017a; Li et al., 2017b; Sulayman et al., 2018; Gong et al., 2019). With the advent of SNP array genotyping technologies, genome-wide association studies (GWAS) using bi-allelic markers have become feasible, and they have been performed in the ovine species to investigate, among others, wool traits (Wang et al., 2014; Bolormaa et al., 2017). An additional approach for connecting DNA to phenotype is the detection of evidence of selective pressure in specific genomic regions by using genome-wide SNP genotype data, also referred to as "selection signatures" analysis. This method has emerged mainly because (i) it does not require the use of phenotypic records, and (ii) unlike GWAS, it can detect selection signatures also when anthropogenic selection has determined complete fixation of the favorable allele (e.g., Qanbari et al., 2014). These features are both relevant in studies addressing genotype–phenotype associations for wool traits, where availability of phenotypic records may represent a limiting issue, and long-term intensive human selection toward wool attributes is likely to have been responsible for the complete prevalence, in the selected populations, of the desired allele. In some of the studies where selection signatures for wool traits have been described, identification of regions affecting wool attributes was not the unique or major goal, with repercussions of this conceptual set-up on the choice of breeds to be contrasted (Zhang et al., 2013; Fariello et al., 2014; Wang et al., 2015; Wei et al., 2015; Seroussi et al., 2017; Rochus et al., 2018). To our knowledge, only two analyses of selection signatures specifically targeting wool attributes have been performed so far (Demars et al., 2017; Gutierrez-Gil et al., 2017). Out of them, only the latter was centered on Merino sheep, which were contrasted, in that study, to the coarse-wool Churra sheep from Spain. Our study follows up on the work by Gutierrez-Gil et al. (2017) to further investigate selection signatures in various Merino (and Merino-derived) sheep breeds under different scenarios, using two different approaches: a classical F_{ST}-outlier method and a less usual one based on the analysis of local ancestry in admixed populations (Sankararaman et al., 2008). Signals observed using both methods are also compared with genome-wide patterns of distribution of ROH (runs of homozygosity) islands. We specifically adopted here a multi-cohort approach with the aim of retaining only the most reliable signals.

MATERIALS AND METHODS

Genotypic Data

A total of 459 unrelated animals arranged in 11 breeds were used in this study (**Table S1**). Out of them, six were Merino or Merino-derived breeds (Spanish Merino, Australian Merino, Rambouillet, Gentile di Puglia, Sopravissana, and Chinese Merino), and five had no known Merino background (Churra, Ojalada, Bergamasca, Appenninica, and Tibetan) and belonged to the category of "coarse wool" sheep, not purposely selected for wool quality traits (**Data Sheet 1**). SNP genotypes had been generated in previous published studies (**Table S1**) by using the Illumina OvineSNP50 Genotyping BeadChip. The whole SNP genotype dataset is available on the WIDDE database (http://widde.toulouse.inra.fr/widde/). The following quality control criteria were applied: (i) individuals with genotyping rate ≤ 90% (command *–mind 0.1*) were removed; (ii) loci with call rate ≤ 99% (command *–geno 0.01*), minor allele frequency ≤ 0.005 (command *–maf 0.005*), and non-autosomal loci were removed; and SNP positions were updated according to the sheep map version Oar_V4. All the above procedures were performed using the PLINK software v. 1.07 (Purcell et al., 2007).

Inference of Local and Global Merino Ancestry

We used the LAMP (Local Ancestry in adMixed Populations) software (Sankararaman et al., 2008) to estimate the individual's

local ancestry of Merino proportion under various scenarios, each contrasting a Merino *versus* a non-Merino breed, for a total of four cohorts (**Table S2**). LAMP is a method for estimating ancestries at each locus in a population of admixed individuals i.e., populations formed by the mixing of two or more ancestral populations. The software operates on sliding windows of contiguous SNPs and assigns ancestries by combining the results with a majority vote. The following default settings were adopted: number of generations since admixture (g) = 7 and recombination rate (r) = 1E−08. We opted for adopting default settings in all the tested scenarios since (i) the method was shown to provide robust estimates under different setting configurations in both the literature (Sankararaman et al., 2008) and our preliminary analyses (data not shown). The fraction of global admixture (α) was determined, for each scenario, using the ADMIXTURE software (Alexander et al., 2009). We ran LAMP in the LAMPANC mode, i.e., providing allele frequencies of the two ancestral population proxies. The LAMP analysis provides, among other output results, the marker average ancestry (MAA) related to the two considered ancestral populations. Only MAAs representing the Merino fraction were considered in this study to identify the significant region supposed to be under selection. To this aim, both of the following criteria should be respected by the putative selection signature: (i) local Merino MAA higher than the genome-wide Merino MAA and (ii) being included in the top 5% of SNPs ranked by MAA of Merino proportion.

Detection of F_{ST}-Outlier Markers

We adopted the F_{ST}-outlier approach implemented in BayeScan (Foll and Gaggiotti, 2008) to detect markers putatively under differential selection pressure in Merino and non-Merino sheep breeds, respectively. To this aim, we performed nine pair-wise comparisons, contrasting each time a Merino *versus* a non-Merino sheep breed (**Table S3**). For each cohort, loci that displayed q-val < 0.05 were retained as putatively under selection. Next, we looked for loci that resulted to be putatively under selection in at least four pair-wise comparisons out of nine. For each SNP satisfying the above criteria, we then moved upstream and downstream its position, looking for additional loci with q-val < 0.05 in at least a single pair-wise comparison, and located within 200 kb intervals. We repeated the above process until the next SNP with q-val < 0.05 in at least a single pair-wise comparison was located at a distance higher than 200 kb. Finally, we defined the regions putatively under selection based on the position of the first and the last of the SNPs satisfying the above criteria.

Runs of Homozygosity

ROH were estimated for each animal belonging to the considered breeds using PLINK (Purcell et al., 2007). The following criteria were used to define the ROH: (i) no missing SNP and no heterozygous genotype were allowed in the ROH, (ii) the minimum number of SNPs that constituted the ROH was set to 25, (iii) the minimum SNP density per ROH was set to one SNP every 100 kb, and (iv) the maximum gap between consecutive homozygous SNPs was 250 kb. The minimum length that constituted the ROH was set to 500 Mb. To identify the genomic regions of high homozygosity, the amount of times that each SNP appeared in the ROH was considered and normalized by dividing it by the number of animals included in the analysis. To identify the genomic regions of "high homozygosity," also called ROH islands, the top 0.999 SNPs of the percentile distribution of the locus homozygosity range within each breed were selected.

Gene and QTL Content of Regions Identified as Under Selection

Annotated genes within the genomic regions putatively under selection were obtained from https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=gene&acc=101104604 (NCBI Sheep Genome Data Viewer). The Sheep QTL Database, available at https://www.animalgenome.org/cgi-bin/QTLdb/OA/srchloc?chrom=19&qrange=454178-607539&submit=GO, was interrogated for the presence of QTLs (quantitative trait loci) and significant association signals in the genomic regions identified in this study as putatively under selection. To investigate the biological function and the phenotypes that are known to be affected by each annotated gene, we conducted a comprehensive search in the available literature and public databases, such as NCBI (https://www.ncbi.nlm.nih.gov/), GeneCards (https://www.genecards.org), UniProt (www.uniprot.org), and Amigo2 Gene Ontology database (http://amigo.geneontology.org/amigo). Furthermore, we performed a gene network analysis by using GeneMANIA (Mostafavi et al., 2008). This tool allows to build weighted interaction networks using as a source a very large set of functional association data including protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity (see http://pages.genemania.org/help/for a more detailed description of the considered network categories).

RESULTS

Signals of Selection Detected *via* "Local Ancestry"

Preliminary to the local ancestry analysis, we performed a "global ancestry" analysis using the Bayesian approach implemented in the software ADMIXTURE (Alexander et al., 2009). Individual proportions of global admixture (α) are presented, for the four considered breeds, in **Figure S1**. The observed patterns support, for all the tested breeds, the formulated scenarios, i.e., that each breed could be considered to derive from the crossbreeding of a given Merino and a given non-Merino breed (breed A and breed B, respectively, in **Table S2**). Putatively selected regions, identified from LAMP results, are shown, for the four considered breeds, in **Table S4–S7**. An excess of Merino ancestry was observed at 26, 24, 17, and 22 regions for Australian Merino, Chinese Merino, Sopravissana, and Spanish Merino, respectively. A number of regions were shared by at least three out of the four breeds (**Table 1**) and, among them, two large regions, on OAR 17 (overlapping signals at 48,474,658–58,410,640 bp) and OAR 18 (overlapping signals at 42,864,163–51,943,741 bp), were shared by all the four breeds.

TABLE 1 | Signals of selection detected *via* local ancestry, shared by at least three of the four considered breeds.

OAR	Australian Merino				Chinese Merino			Sopravissana			Spanish Merino			
	SNP ID	Position (bp)	MAA		SNP ID	Position (bp)	MAA	SNP ID	Position (bp)	MAA	SNP ID	Position (bp)	MAA	
5	Start rs418698529	26341551	1	Start	rs414589041	56778930	0.97	Start rs409779782	34321292	0.89				
	End rs414589041	56778930	1	End	rs415945949	66146282	0.97	End rs407964514	56019746	0.89				
13	Start rs421927509	62007588	1	Start	rs160614980	67099575	0.97				Start rs160604475	63340057	0.92	
	End rs421743434	82928768	1	End	rs421743434	82928768	0.97				End rs421743434	82928768	0.92	
14	Start rs405740814	173039	1	Start	rs403113459	6678882	0.97				Start rs405740814	173039	0.80	
	End rs409789065	18615310	1	End	rs423800989	12841601	0.97				End rs407274307	22608758	0.80	
15	Start rs398486856	747553	0.97					Start rs430560325	24611939	0.87	Start rs398442436	19097698	0.96	
	End rs417609233	48071326	0.97					End rs410686010	42463484	0.87	End rs404810360	39888859	0.96	
16	Start rs429048373	41142879	0.93					Start rs419085314	41237027	0.91	Start rs403615656	31260763	0.92	
	End rs399764897	49217439	0.93					End rs421054947	55127918	0.91	End rs408523982	54928813	0.92	
17	*Start rs398968259*	*20314798*	*0.97*	*Start*	*rs424790285*	*47829307*	*0.95*	*Start rs419855399*	*46130137*	*0.91*	*Start rs426738329*	*48474658*	*0.88*	
	End rs414015395	*71582708*	*0.97*	*End*	*rs400781870*	*72084885*	*0.95*	*End rs411845108*	*58410640*	*0.91*	*End rs414015395*	*71582708*	*0.88*	
18	*Start rs416601769*	*42864163*	*0.97*	*Start*	*rs414605156*	*33332554*	*1*	*Start rs401597794*	*28792875*	*0.83*	*Start rs415215918*	*42698731*	*0.88*	
	End rs416850669	*51943741*	*0.97*	*End*	*rs400436533*	*59172019*	*1*	*End rs415998758*	*68381958*	*0.83*	*End rs415998758*	*68381958*	*0.88*	
19	Start rs409364012	28430094	0.97	Start	rs419333175	31614145	0.97				Start rs408609148	34392185	0.84	
	End rs412219091	53122560	0.97	End	rs412185520	60327629	0.97				End rs412185520	60327629	0.84	

OAR, sheep chromosome. SNP ID, name of the SNP locus. MAA, marker average ancestry (see the section Materials and methods). In italics, regions that were shared by all the four breeds.

Signals of Selection Detected *via* the "F_{ST}-Outlier" Method

Results of the analyses performed using the F_{ST}-outlier approach implemented in BayeScan are presented, for the nine pair-wise comparisons involving Merino and non-Merino breeds, in **Table S8**. The highest number of significant SNPs (277) was detected for the contrast Chinese Merino *vs*. Tibetan. A summary of the obtained results is presented in **Table 2**. Four regions putatively under differential selection pressure were identified, on OAR3 (15,382,6281–154,318,689 bp), OAR10 (29,392,142–29,776,019 bp), OAR13 (62,707,138–62,747,155 bp), and OAR19 (454,178–607,539 bp). Interestingly, in the region on OAR13, one SNP (rs415003205) had q-val < 0.05 in all the nine considered pair-wise comparisons. This is a deep intronic variant (G/A) located at 5,264 bp downstream the end of the first exon of the RALY gene. For the region on OAR3, the locus displaying the highest number of pair-wise comparisons showing signal of selection (six out of nine) was rs423370130 (154,072,493 bp). For the region on OAR19, the highest number of pair-wise comparisons showing signal of selection was five (rs404730996). The large region on OAR10 displayed a maximum of six pair-wise comparisons showing signal of selection, at 29,413,536 bp (rs401979890). In the same region, two loci, out of which one (rs414794714, at 29776019 bp) rather far from rs401979890, had five pair-wise comparisons showing signal of selection. This pattern suggests that the considered region may harbor two different selection signatures.

The loci that provided overlapping signals for the same Merino (or Merino-derived) breed with both the "local ancestry" and the "F_{ST}-outlier" methods are highlighted in **Table S8**, while in **Tables S4** to **S7**, the putatively selected regions, detected using the "local ancestry" method, where at least one significant SNP in at least one pair-wise comparison of the "F_{ST}-outlier" method involving the corresponding Merino (or Merino-derived) breed was observed, are highlighted in light yellow. In general, the majority of the regions (16/26, Australian Merino; 15/24, Chinese Merino; 9/17 Sopravissana; 12/22 Spanish Merino) showed overlapping signals between the two methods.

Gene and QTL Content of Putatively Selected Regions

The two large regions on OAR 17 and 18, detected as putatively selected by "local ancestry" analysis, contain 148 and 70 genes, respectively (**Table S9** and **S10**). These regions were screened for the presence of known QTLs in sheep (**Table S11A**). Interestingly, we found two QTLs associated with a wool trait, notably "greasy fleece weight," at positions 49,606,819–49,606,859 bp and 51,061,367–51,061,407 bp, respectively, in OAR17 (Ebrahimi et al., 2017), and one QTL associated with "staple length," at position 9,943,363–68,604,602 bp in OAR18 (Allain et al., 2006). The OAR17 QTL at position 49,606,819–49,606,859 bp is located in an inter-genic region, between *LOC101120019* (60S ribosomal protein L10a-like, at position 49,486,725–49,520,271 bp) and *TMEM132B* (transmembrane protein 132B, at position 49,844,529–50,246,808 bp) (data not shown). Similarly, the OAR17 QTL at position 51,061,367–51,061,407 bp is located

TABLE 2 | Summary results of the F_{ST}-outlier approach for the nine pair-wise comparisons between Merino and non-Merino breeds.

Loci			Pair-wise comparisons									N
OAR	**SNP ID**	**Position**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
3	rs429917763	153826281						■		■		2
3	rs426111530	153889169						■		■		2
3	rs408016275	153927239						■	■	■		3
3	rs414901427	153976304						■	■	■		3
3	rs409568101	153996225									■	1
3	rs416115321	154033734	■								■	2
3	rs423370130	154072493	■		■		■	■	■	■		6
3	rs417916710	154223123						■		■		2
3	rs159858948	154318689	■					■		■		3
10	rs419203432	29392142				■		■		■		3
10	rs401979890	29413536	■	■	■		■	■	■	■		6
10	rs413264476	29453722										1
10	rs424871667	29479711	■	■	■		■	■		■		5
10	rs399348601	29489616					■		■		■	3
10	rs425859016	29660838									■	1
10	rs404720287	29685665								■	■	2
10	rs415997827	29742016					■	■	■	■		2
10	rs414794714	29776019	■				■	■	■	■		5
13	rs401457425	62707138	■		■			■		■		4
13	rs415003205	62747155	■	■	■	■	■	■	■	■	■	9
19	rs421064536	454178						■				1
19	rs409839516	504608					■	■				2
19	rs404730996	566456			■		■	■	■	■		5
19	rs424406294	607539					■					1

OAR, sheep chromosome. SNP ID, name of the SNP locus. N, number of pair-wise comparisons showing signal of selection for the considered SNP. 1, Australian Merino vs. Churra. 2, Australian Merino vs. Ojalada. 3, Spanish Merino vs. Churra. 4, Spanish Merino vs. Ojalada. 5, Gentile di Puglia vs. Appenninica. 6, Gentile di Puglia vs. Bergamasca. 7, Sopravissana vs. Appenninica. 8, Sopravissana vs. Bergamasca. 9, Chinese Merino vs. Tibetan.

in an inter-genic region, between *LOC101115905* (refilin-A, alias "family with sequence similarity 101, member A" or "filamin-interacting protein FAM101A," at position 51,021,418–51,035,210 bp) and *LOC106991703* (a long non-coding RNA, at position 5,118,8762–51,255,313 bp) (data not shown). The large OAR18 QTL at position 9,943,363–68,604,602 bp includes 693 genes (data not shown) and was hence not useful to refine the signal position. Therefore, the 70 genes detected in the putatively selected region on OAR18 were screened for inclusion in the output of the human GeneCards database using the queries "hair," "wool," and "horn," selected as the most representative of the Merino phenotype. While none of the genes was retrieved when using "wool" or "horn" keywords, a total of 16 out of 70 (23%) genes were retrieved when using the keyword "hair" (**Table S10**). Notably, three genes displayed particularly high GeneCards relevance scores: *NFKBIA* (16.1), *SEC23A* (15.27), and *PAX9* (7.17).

The four regions detected as putatively selected by "F_{ST}-outlier" analysis contain five (OAR3), four (OAR10), one (OAR13), and two (OAR19) genes (**Tables S12A–D**). Also, these regions were screened for the presence of known QTLs in sheep (**Table S11B**). On OAR3, a QTL associated with wool traits (notably, "staple length") had been previously mapped, within a large chromosome interval (region 1,184,337–224,283,230 bp) encompassing the region detected in this study (Ponz et al., 2001). On OAR10, a genome-wide association study for wool traits in Chinese Merino detected a significant SNP for fiber diameter at position 30 Mb, and several SNPs significant for crimp at 26–27 Mb (Wang et al., 2014). On OAR13, one SNP at 62.9 Mb was associated with wool fiber diameter (Bolormaa et al., 2017).

The results of the gene network analysis for the genes located in the putatively selected regions mentioned above are presented in **Figure 1** and **Table S13**. A total of 57 links are reported for the considered 29 genes, out of which 9 genes had been detected in this study as putatively selected. Interestingly, links were observed not only between genes detected as putatively selected using the same approach, either LAMP or the F_{ST}-outlier, but also between genes detected as putatively selected using different approaches (SEC23A/MSRB3, RXFP2/PAX9, EIF2S2/TMEM132B, SEC23A/RXFP2), thus suggesting their complementarity in selection signature detection.

Runs of Homozygosity

Several genomic regions that frequently appeared in a ROH were identified within each breed. **Table 3** provides the chromosome position, and the start and end of the detected ROH islands. The top 0.999 SNPs of the percentile distribution of locus homozygosity values led to the use of different thresholds for each breed (from 0.166, in Gentile di Puglia, to 0.261, in

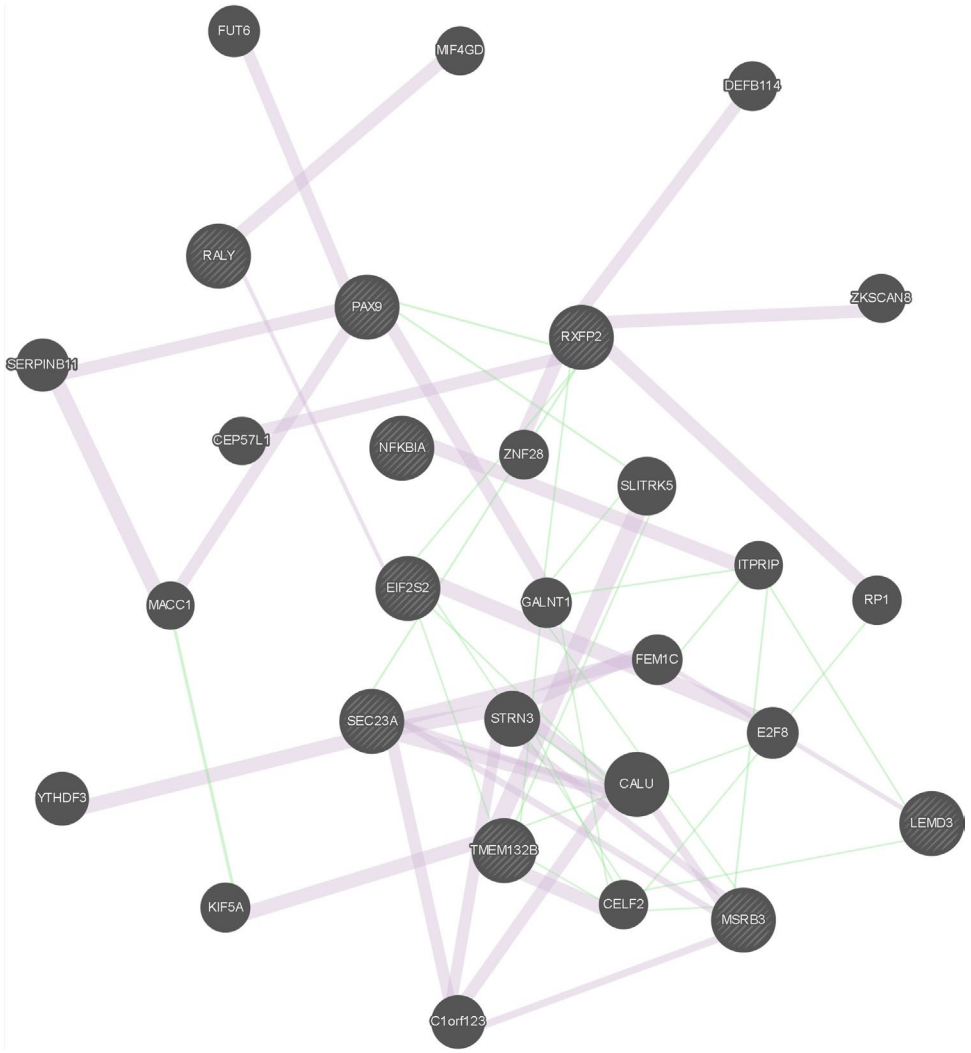


FIGURE 1 | Graphical representation of the gene network analysis. Query genes, i.e., genes detected in this study as putatively selected, are indicated with stripes. Links in light purple indicate the network category co-expression (two genes are linked if their expression levels are similar across conditions in a gene expression study. Most of these data are collected from GEO, the Gene Expression Omnibus); only data associated with a publication are collected. Links in green indicate the network category genetic interaction. (Two genes are functionally associated if the effects of perturbing one gene were found to be modified by perturbations to a second gene. These data are collected from primary studies and BioGRID).

Chinese Merino). The genomic distribution of ROH islands was clearly non-uniform among breeds. Gentile di Puglia showed the highest number (21) of ROH islands, followed by the Spanish Merino (12). Gentile di Puglia, together with Sopravissana, also displayed large proportions (33.3% and 50%, respectively) of ROH islands longer than 5 Mb. These results may well reflect the serious bottlenecks experienced by these breeds in the last 70 years. Three overlapping ROH islands were observed between breed pairs. Spanish Merinos and Gentile di Puglia breeds showed a common 5 Mb genomic region on OAR12 (47,013,871 to 52,019,776 bp). Smaller (<1 Mb) genomic regions were shared between Gentile di Puglia and Chinese Merino on OAR2 (99,442,430 to 100,215,565 bp) and between Chinese Merino and Appenninica on OAR16 (30,100,068 to 30,670,323 bp).

When comparing the ROH islands observed within each Merino (or Merino-derived) breed (**Table 3**) with regions detected by "local ancestry" analysis involving the same Merino (or Merino-derived) breed (**Tables S4–S7**), we found little overlapping. In particular, the two ROH islands detected on OAR25 in Australian Merino were both included in the LAMP region detected for the same breed on the same chromosome. Similarly, the four ROH islands detected on OAR6 in Spanish Merino were all included in the LAMP region detected for the same breed on the same chromosome. No overlapping was observed for Chinese Merino and Sopravissana.

When comparing the ROH islands observed within each Merino (or Merino-derived) breed (**Table 3**) with SNPs detected as significantly differentiated in "F_{ST}-outlier" pair-wise contrasts involving the same Merino (or Merino-derived) breed

TABLE 3 | List of genomic regions of extended homozygosity (ROH islands) identified in the considered Merino and non-Merino breeds.

Breed (locus homozygosity threshold)	OAR	Start	End	N.
Australian Merino (0.196)	3	33,232,651	34,050,238	19
	25	19,861,459	20,568,885	15
	25	21,904,797	22,347,925	13
Spanish Merino (0.231)	1	249,023,519	249,191,465	5
	6	32,912,993	35,003,625	47
	6	37,126,564	38,480,285	30
	6	39,589,194	39,715,842	4
	6	40,342,592	43,655,868	72
	7	1,830,665	3,913,607	45
	12	31,598,245	34,784,182	72
	12	38,121,281	38,765,181	15
	12	41,659,697	42,066,590	8
	12	47,013,871	52,019,776	93
	12	53,371,161	56,327,304	61
	12	63,599,219	64,794,499	26
Gentile di Puglia (0.166)	1	211,018,133	216,875,491	107
	1	270,012,629	271,410,339	28
	2	99,442,340	101,718,337	110
	2	202,780,179	203,472,364	20
	2	217,829,936	223,681,332	116
	2	223,981,060	226,600,106	55
	2	240,008,834	243,300,137	61
	3	211,410,359	218,603,858	139
	5	93,955	3,046,488	66
	5	25,728,102	28,636,862	65
	10	13,800,857	13,988,776	5
	10	14,165,558	20,000,839	111
	10	20,273,388	23,396,844	64
	12	44,162,620	52,019,776	154
	12	70,838,617	78,861,071	151
	17	8,532,536	9,566,661	21
	17	17,289,600	17,844,323	10
	18	3,363,915	6,142,721	47
	22	11,697,681	12,751,792	20
	26	8,124,065	13,498,474	91
	26	17,062,097	19,422,382	38
Sopravissana (0.208)	5	65,184,537	69,108,780	79
	5	72,592,808	73,484,347	20
	5	73,814,249	81,172,608	150
	15	17,158,900	22,517,143	94
	22	18,932,514	24,872,911	109
	22	28,773,373	30,395,735	29
Chinese Merino (0.261)	2	92,669,379	95,401,516	60
	2	95,689,756	100,215,565	82
	3	142,710,943	142,862,611	3
	6	30,411,203	30,508,550	5
	10	67,762,612	70,157,217	45
	16	30,100,068	30,670,323	16
Churra (0.229)	8	32,122,858	34,554,414	49
Ojalada (0.167)	21	17,001,944	19,824,196	44
Bergamasca (0.167)	2	10,823,174	12,893,239	45
	9	36,932,939	37,952,215	24
Appenninica (0.208)	4	44,524,519		1
	16	26,703,405	30,695,539	88
Tibetan (0.243)	8	26,853,601	30,314,835	77
	23	77,105,889	83,429,082	103
	23	90,336,146	94,196,356	64
	23	98,904,521	101,871,791	52
	23	104,351,418	108,119,862	57
	23	114,615,110	120,163,589	90
	23	121,125,567	125,963,555	81

OAR, sheep chromosome. N, number of SNPs.

(**Table S8**), some overlapping was observed. In particular, the SNP rs408794746 (34,050,238 bp in OAR3) was significantly differentiated when contrasting Australian Merino with Churra and Ojalada and was also detected within a ROH island in Australian Merino. The significantly differentiated SNP rs425817109 (34,390,603 bp) in the Spanish Merino *vs*. Churra comparison, and the SNP rs400309388 (34,699,452 bp) in the Spanish Merino *vs*. Ojalada comparison, were both included within a ROH island detected in Spanish Merino (OAR12). The SNP rs403786137 (215,181,085 bp in OAR1) in the Gentile di Puglia *vs*. Bergamasca comparison was included within a ROH island detected in Gentile di Puglia. The SNP rs398231484 (216,225,845 bp in OAR3) in the Gentile di Puglia *vs*. Appenninica comparison was included within a ROH island detected in Gentile di Puglia. The SNP rs407100968 (45,671,005 bp) in the Gentile di Puglia *vs*. Appenninica comparison and the SNP rs417849493 (48,709,065 bp) in the Gentile di Puglia *vs*. Bergamasca comparison were both included within a ROH island detected in Gentile di Puglia (OAR12). The SNP rs399908187 (68,477,988 bp in OAR5) in the Sopravissana *vs*. Bergamasca comparison was included within a ROH island detected in Sopravissana. No significantly differentiated SNP overlapping with ROH islands was detected for the Chinese Merino breed.

DISCUSSION

Comparison Among Approaches for Selection Signatures Detection

In this study, we investigated selection signatures in various Merino (and Merino-derived) sheep breeds using two different approaches. While the "F_{ST}-outlier" is considered a classical method for identification of regions putatively under differential selection in pairs of breeds (or group of breeds), the analysis of local ancestry, i.e., the genetic ancestry of an individual at a particular chromosomal location, in admixed populations to detect genomic regions where a significant excess of ancestry from a given parental breed exists (also known as "admixture mapping") is so far a less popular approach. Among the "admixture mapping" approaches, LAMP has some interesting features that prompted us to opt for this method. Unlike algorithms that are based on reference

haplotype frequencies for each of the parental populations, for which larger sample sizes are required to capture haplotypic diversity, LAMP relies on reference allele frequencies (Shriner, 2013) and is consequently less affected by a reduced sample size. Also, it operates on sliding windows of contiguous SNPs, using a "majority vote" for each locus, over all windows that overlap with the SNP, in order to decide the most likely ancestral population at the marker. This simple approach has been shown to provide fast and robust estimates (Sankararaman et al., 2008). Despite LAMP has been developed for estimation of the locus-specific ancestry in recently admixed populations, it has been shown to be robust to inaccuracies in the parameter "number of generations since the admixture." A critical issue, limiting the widespread use of LAMP, is represented by the choice of the external reference samples to be used as proxies for the true ancestral populations, as the latter are generally not available for sampling (Shriner, 2013). In this study, a set of four hypotheses, each including a test breed and two proxies for the parental populations, were formulated based on historical knowledge on the origin of breeds and the inferred proportions of global admixture. These have to be interpreted with caution given the possible influence of complex patterns of historical admixture known among Merino and Merino-derived sheep populations (Ciani et al., 2015). The four hypotheses were hence tested using the algorithm implemented in LAMP. On the other side, for the "F_{ST}-outlier" approach, we were able to define, a set of nine pair-wise comparisons by contrasting (i) Merino populations of Iberian origin (Spanish Merino and Australian Merino, respectively) with non-Merino populations of Iberian origin (Churra and Ojalada, respectively), (ii) Merino populations of Italian origin (Gentile di Puglia and Sopravissana, respectively) with non-Merino populations of Italian origin (Appenninica and Bergamasca, respectively), and (iii) a Merino population of Asian origin (Chinese Merino) with a non-Merino population of Asian origin (Tibetan). The rationale behind the above pairing is that differentially selected loci may be easier to detect when contrasting more homogeneous breeds, such as Merino *versus* non-Merino breeds from the closest geographical area (Manunza et al., 2016).

Consistently with expectations, the two adopted approaches produced only partly overlapping signals. Indeed, the two methods rely on different algorithms and different assumptions, which also imposed a different organization of the dataset used with the two approaches (four single-breed tests *vs.* nine pair-wise comparisons, for the "local ancestry" and the FST-outlier" approaches, respectively), thus hampering direct head-to-head comparisons. Notwithstanding, the majority of the regions detected using the "local ancestry" method showed overlapping signals with the "F_{ST}-outlier" results. Although we interpreted the above as evidence supporting the robustness of the obtained results, it must be taken into consideration that regions identified by "local ancestry" were generally large, and significant SNPs detected *via* the "F_{ST}-outlier" method may likely occur in there by chance. Indeed, the number of loci identified as putatively under selection pressure using the "F_{ST}-outlier" method largely exceeds the number of putatively selected regions identified using the "local ancestry" approach.

In this study, we also investigated genomic regions of high homozygosity (ROH islands), as these have been shown to be abundant in regions under positive selection (Metzger et al., 2015; Mastrangelo et al., 2017; Purfield et al., 2017; Talenti et al., 2017a; Talenti et al., 2017b; Mastrangelo et al., 2018). While we observed little overlapping between ROH islands and regions identified *via* "local ancestry," some overlapping was observed between ROH islands and SNPs detected as significant using the "F_{ST}-outlier" approach. ROH islands may be the consequence of the genetic hitchhiking phenomenon at loci physically linked to the variant site under direct positive selection pressure. The "local ancestry" approach looks for regions with an excess of ancestry from one of the two parental populations, and not necessarily these regions have to display high homozygosity, although this feature is likely to be observed in case of a strong selective sweep. Similarly, in the "F_{ST}-outlier" approach, homozygous genotypes (for different alleles in the two breeds) at loci physically linked to the variant site under direct positive selection pressure may display high frequencies if a strong differential selection existed among the two considered breeds. Also, the argumentation reported above may apply here: the larger number of loci identified as putatively under selection pressure using the "F_{ST}-outlier" method may be more likely to occur by chance within large ROH islands compared to the fewer genomic regions identified *via* "local ancestry." Moreover, ROH analysis might detect selection related to any trait, while contrasting Merino and non-Merino is more likely to detect signals related to this specific trait. Finally, the existence of ROH islands may be due to non-genetic factors such as demography.

Best Candidate Regions and Putatively Selected Genes

As the aim of this study was to identify loci most likely associated with the Merino phenotype, we arbitrarily identified (i) the best candidate regions detected *via* the "local ancestry" approach as those being shared by all the four breeds and (ii) the best candidate SNPs detected *via* the "F_{ST}-outlier" approach as those observed in at least 70% of the pair-wise comparisons (six out of nine). Based on the above, two large regions on OAR17 and OAR18 were retained for (i), and three, on OAR3, OAR10, and OAR13, for (ii). The robustness of the adopted procedure was also suggested by the occurrence, in all of the five regions, of QTLs/associations known to be related to wool traits in the ovine species. Moreover, at OAR17, combining analysis of "local ancestry" and inspection of the sheep QTL database allowed to significantly shorten the candidate interval. On the contrary, known QTLs for wool traits described on OAR18 and OAR3 are mapped within extremely large chromosome intervals. These were identified by Allain et al. (2006) and Ponz et al. (2001) who performed whole-genome scans using microsatellite markers on experimental flocks obtained crossing Lacaune with Sarda, and Berrichon du Cher (a Merino-derived breed) with Romanov (a non-Merino breed), respectively. In what follows, the gene content of the best candidate regions is presented by chromosome order.

OAR3

The region detected on OAR3 contains five genes, *LOC105609945* (long noncoding RNA), *MSRB3* (methionine sulfoxide reductase B3), *LOC105609947* (long noncoding RNA), *LOC105609948* (a pseudo-gene), and *LEMD3* (LEM domain containing 3). Interestingly, the sub-region containing the genes *MSRB3*, *LOC105609947*, and *LEMD3* was found to harbor a selection signature putatively for tail fat deposition in previous studies contrasting thin- *vs.* fat-tail sheep breeds, from China (Yuan et al., 2017), and from North Africa and the Chios island (Mastrangelo et al., 2019), for adaptation when contrasting the Red Maasai sheep with the Ethiopian Menz (Fariello et al., 2014), and for ear morphology in Chinese sheep breeds (Wei et al., 2015) and in French Suffolk sheep (Rochus et al., 2018). The latter suggest to consider the genes encoded by the signal on OAR3, notably MSRB3 and LEMD3, as candidates for ear size based on literature showing the possible role of the two genes in ear position in dogs (Vaysse et al., 2011) and ear size in pigs (Wilkinson et al., 2013). A more detailed discussion of each single gene in the OAR3 selection signature is provided below.

LOC105609945—No evidence for involvement of *LOC105609945* in any peculiar Merino feature was found.

MSRB3—The methionine sulfoxide reductase B3 (*MSRB3*, alias *DFNB74*) catalyzes the reduction of free and protein-bound methionine sulfoxide to methionine. This antioxidant repair enzyme has been described in human epidermal keratinocytes and melanocytes, as well as in hair follicles (Taungjaruwinai et al., 2009). It has been shown to be expressed also in inner and outer hair cells of mouse inner ear (Ahmed et al., 2011). Diseases associated with *MSRB3* include deafness (https://www.genecards.org). Down-regulation of *MSRB3* has been shown to impair the normal auditory system development through hair cell apoptosis in zebrafish (Shen et al., 2015). The gene has been found in previous selection signatures studies in sheep (Kijas et al., 2012; Fariello et al., 2014; Wei et al., 2015; Manunza et al., 2016; Yuan et al., 2017; Rochus et al., 2018; Mastrangelo et al., 2019). More interestingly, for this study, it has been found within a selection signature observed contrasting fine-wool Merino and coarse-wool Churra sheep breeds (Gutiérrez-Gil et al., 2017). Another line of evidence for the involvement of *MSRB3* in hair/wool physiology comes from the observation that actin's polymerization properties and actin cytoskeletal-mediated events, such as correct bristle development, which are altered by specific oxidation of its conserved methionine (Met)-44 residue on the pointed-end of actin subunits, are rescued by a methionine sulfoxide enzyme reductase (SelR/MsrB) in *Drosophila (*Hung et al., 2013). In this species, actin plays a role not only in bristle but also in wing hair development (Guild et al., 2005). In mammals, actin has been shown to be one of the major components of both the water-soluble and -insoluble fraction from hair and hair follicles (Vermorken et al., 1981; Lee et al., 2006). Actin bundles in the hair follicle would act as stress fibers and serve as a tensile scaffold for the growth and integrity of the hair follicle (Furumura and Ishikawa, 1996). In Tibetan sheep, microRNAs differentially expressed in wool follicles during anagen, catagen, and telogen phases, thus potentially regulating wool follicle development, targeted, among others, genes in the pathways that regulate the actin cytoskeleton (Liu et al., 2013). *MSRB3* also contained the (intronic) SNP that, in this study, displayed the highest number of "F_{ST}-outlier" pair-wise comparisons showing signal of selection observed in the OAR3 region.

LOC105609947—No evidence for involvement of *LOC105609945* in any peculiar Merino feature was found.

LOC105609948—It's a pseudo for the ubiquitin-conjugating enzyme E2 D3 gene (*UBE2D3*), which is part of the bone morphogenic protein (BMP) signaling pathways (gene ontology database accession ID: GO:0030509). BMP ligands (*BMP2* and *BMP4*) when expressed in dermal macro-environment during telogen (resting phase of hair cycle) have been shown to strongly suppress ability of resting hair follicles to be reactivated and grow again (International Patent no. WO2010059861A1 available at https://patents.google.com/patent/WO2010059861A1).

LEMD3—As previously mentioned, the LEM domain containing three gene (alias *MAN1*) has been found in various selection signatures studies in sheep (Fariello et al., 2014; Wei et al., 2015; Manunza et al., 2016; Yuan et al., 2017; Rochus et al., 2018; Mastrangelo et al., 2019), including the study by Gutierrez-Gil et al. (2017) where fine-wool Merino and coarse-wool Churra sheep breeds were contrasted. Moreover, it has been found associated with the abnormal hair quantity phenotype from the HPO Gene-Disease Associations dataset (Köhler et al., 2014).

OAR10

The region detected on OAR10 includes four genes: *LOC106991357* (long noncoding RNA), *LOC101110773* (elongation factor 1-alpha 1-like), *RXFP2* (relaxin/insulin-like family peptide receptor 2), and *LOC106991379* (a pseudo-gene). Despite this region was detected as putatively selected in studies investigating tail fat deposition (Moioli et al., 2015; Yuan et al., 2017; Mastrangelo et al., 2019) and adaptation (Yang et al., 2016; Seroussi et al., 2017), RXFP2 is the most studied gene and is well known for being involved in horn presence/absence and morphology in sheep (Johnston et al., 2011; Kijas et al., 2012; Fariello et al., 2014; Pan et al., 2018). A genome-wide association study for wool traits in Chinese Merino sheep detected, on this chromosome, a significant SNP for "fiber diameter" at position 30 Mb, together with several SNPs significant for "crimp" at 26–27 Mb (Wang et al., 2014). In what follows, a more detailed discussion of the four genes annotated in the OAR10 region is provided.

LOC106991357—No evidence for involvement of this locus in any peculiar Merino feature was found.

LOC101110773—It codes for an elongation factor 1-alpha 1-like. The elongation factor 1-alpha 1 (*EF1A1*) is a GTP-binding protein which has a primary function as an essential house-keeping gene by delivering aminoacyl-tRNAs to the ribosome during the elongation step of protein translation. *EF1A1*, together with genes associated to the Usher syndrome, a congenital disease characterized by perturbation of normal organization and growth of hair bundles within the inner ear, is a downstream target of *GBX2*, which induces *EF1A1* activation upon binding to the *EF1A1* core promoter. *GBX2* has been shown to be expressed in the otic placode, which develops into the inner ear. Loss-of-function and mis-expression studies have shown that *GBX2*

is essential for development of the inner ear sensory organs. However, neither direct evidence for involvement of *EF1A1* in hair bundles organization and growth nor in any peculiar Merino phenotype has been found so far. Another elongation factor type (*EF1Bγ*) has been proposed to bind to keratin (Kim et al., 2006). The presence of large amounts of *EF1Bγ* in keratin bundle rich hair fibers would support its biological role in the intermediate filament organization (Sasikumar et al., 2012).

RXFP2—This gene is involved, among others, in the biological process "activation of adenylate cyclase activity" (https://www.uniprot.org/uniprot/Q8WXD0). Adenylate cyclase is responsible for the synthesis of 3',5'-cyclic adenosine monophosphate (cAMP). Agents that increase cAMP levels have been shown to be potent inhibitors of human and mouse hair follicle growth (Harmon and Nevins, 1997). However, we cannot exclude that, in our tested scenarios, different alleles at this gene may have been differentially selected in the considered breeds as a consequence of selection toward different horn phenotypes. Rochus et al. (2018) highlighted that a number of single nucleotide polymorphisms exist in French sheep in the region extending 100 Kb upstream of *RXFP2*, with haplotypes in polled sheep being distinct from those observed in horned sheep. From these findings, they suggested that multiple ancient mutations, rather than a single mutation, are likely affecting horn phenotypes. Pan et al. (2018) reported strong association between three SNPs within the *RXFP2* gene and horn sizes in a Tibetan population characterized by the presence of animals with heterogeneous horn types.

LOC106991379—No evidence for involvement of this locus in any peculiar Merino feature was found.

It is worth mentioning that, only 0.076 Mb upstream to LOC106991357 on OAR10, the gene *FRY* is mapped (interval 28,959,450–29,212,913 bp). Looking at our results separately for each tested scenario, we observed that this interval was overlapping with the region detected by the "local ancestry" method in Chinese Merino, as well as with the regions detected by the "F_{ST}-outlier" method in the Chinese Merino *vs.* Tibetan, Australian Merino *vs.* Churra, and Gentile di Puglia *vs.* Appenninica comparisons. Moreover, the *FRY* interval was only slightly upstream to the regions detected by the "F_{ST}-outlier" method in the Australian Merino *vs.* Ojalada (0.2 Mb), Spanish Merino *vs.* Ojalada (0.18 Mb), Gentile di Puglia *vs.* Bergamasca (0.18 Mb), Sopravissana *vs.* Appenninica (0.2 Mb), and Sopravissana *vs.* Bergamasca (0.18 Mb). In sheep, *FRY* has been suggested as a key candidate gene for the piebald phenotype in Merino (Garcia-Gamez et al., 2011) and has been suggested to be associated with the black spot phenotype in Valley-type Tibetan sheep (Wei et al., 2015), and with differences in coat color pigmentation distribution between the Awassi and Afec-Assaf sheep (Seroussi et al., 2017). On the contrary, Zhang et al. (2013) detected *FRY* when contrasting Rambouillet and Suffolk sheep and suggested it to be a candidate gene affecting wool quality. Indeed, *FRY* encodes a protein furry homolog that, in *Drosophila*, has been found in growing hairs (He et al., 2005), and whose disruption has been shown to provoke abnormally branched bristles and strong multiple-hair phenotype, with clusters of epidermal hairs and branched hairs (Cong et al., 2001). Fang et al. (2010), following the transgenic *FRY* protein *in vivo*, found it to be highly mobile and to accumulate at the distal tip of growing bristles and suggest that it could function in directing/mediating the intracellular transport needed for polarized growth.

OAR13

The region detected on OAR13 contains a single gene (*RALY*). It encodes a heterogeneous nuclear ribonucleoprotein that binds poly-U-rich elements within several RNAs and regulates the expression of specific transcripts (Cornella et al., 2017; Rossi et al., 2017). In early 90s, the gene was shown to be involved in the pleiotropic lethal yellow phenotype of the mouse due to a deletion of the genes *RALY* and *EIF2S2* (eukaryotic translation initiation factor 2 subunit 2), upstream the *ASIP* (agouti signaling protein) gene, responsible for the ectopic over-expression of the agouti signaling protein under the control of the *RALY* promoter (Michaud et al., 1993; Duhl et al., 1994). The gene has been repeatedly detected, often together with the neighbor ASIP gene, in association studies concerning skin pigmentation and skin neoplasms, (https://www.ncbi.nlm.nih.gov/gap/phegeni?tab=1&gene=22913; Jacobs et al., 2015), as well as in several other type of cancers, where it is considered to represent a metastatic marker (Roberts et al., 2019). In 2013, a mutation in this gene has been associated with the saddle tan and black-and-tan phenotypes in Basset Hounds and Pembroke Welsh Corgis (Dreger et al., 2013). Similarly, it has been associated with coat color phenotypes in Chinese and Iranian goats (Guo et al., 2018; Nazari- Ghadikolaei et al., 2018). Worth mentioning that a SNP at OAR13 (position 62.9 Mb) was associated by Bolormaa et al. (2017) with wool fiber diameter in Merino sheep. Although the SNP is not far from the *RALY* gene, it mapped within the *EIF2S2* gene, which has been shown to be involved in protection against chemotherapy-induced alopecia (Nasr et al., 2013).

OAR17

Out of the two pairs of genes flanking the two QTLs on OAR17, one (*LOC101120019*) is a pseudo-gene related to a 60S ribosomal protein (L10A), one (*TMEM132B*) codes for a trans-membrane protein, one (*LOC101115905*) codes for refilin-A (alias *FAM101A*), and one (*LOC106991703*) is responsible for the production of a long noncoding RNA. Their possible involvement in the Merino phenotype is discussed here based on evidences from the literature.

LOC101120019—A mutation in a 60S ribosomal protein (L21) has been shown to be involved in hereditary hypotrichosis simplex (HHS), a form of nonsyndromic inherited hair loss disorders (Zhou et al., 2011). 60S ribosomal proteins (L6 and L24) have been shown to be expressed in human anagen hair samples (Carlson et al., 2018). Interestingly, the 60S ribosomal protein L10A has been shown to be expressed in root hairs of *Medicago truncatula* (Covitz et al., 1998). As is common for genes encoding ribosomal proteins, multiple processed pseudo-genes of the 60S ribosomal protein L10A are dispersed through the genome (https://www.ncbi.nlm.nih.gov/gene/4736). In particular, *LOC101120019* on OAR17, being a pseudo-gene, is more likely to play, if any, a regulatory function on the hair physiology.

TMEM132B—The trans-membrane protein 132B is required for normal inner ear hair cell function (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TMEM132E). *TMEM132A*, but not *TMEM132B*, *TMEM132C*, or *TMEM132D*, was found to be expressed in wool follicle bulb of Tibetan sheep during phase transformation from the middle anagen, to catagen and late telogen/early anagen (Liu et al., 2015). *TMEM132E* was found to be highly expressed in murine inner hair cells, and a variant in *TMEM132E* was identified as the most likely cause of autosomal-recessive nonsyndromic hearing loss. Knockdown of the *TMEM132E* ortholog in zebrafish affected the mechanotransduction of hair cells. (Li et al., 2015).

FAM101A—The gene product is involved in the regulation of the perinuclear actin network and nuclear shape through interaction with filamins. It plays an essential role in the formation of cartilaginous skeletal elements (UniProtKB:Q5SVD0). In addition, it has been shown to be differentially expressed in hair follicle stem cells residing in the bulge of mouse hair follicles *versus* the epithelial basal cells outside the bulge (Chang, 2014). *FAM101A* mRNA was detected *via* next-generation sequencing in wool follicle bulb samples of Tibetan sheep from middle anagen, catagen, and late telogen/early anagen phases (Liu et al., 2015). In a genome-wide association study performed using 50 K SNPs in a Baluchi sheep population, one of the significant SNP markers associated with greasy fleece weight was located within *FAM101A* (Ebrahimi et al., 2017).

LOC106991703—No evidence for involvement of *LOC106991703* in any peculiar Merino feature was found.

OAR18

The region identified on OAR18 included a large number of genes (70) that obviously hampered a detailed analysis of the available literature for each single gene. We hence opted for checking which of the above genes could be retrieved by querying the human GeneCards database using keywords representative of Merino phenotypes. While none of the genes was retrieved when using "wool" or "horn" keywords, 16 genes were retrieved when using the keyword "hair" and, among them, three displayed particularly high GeneCards relevance scores (*NFKBIA, SEC23A*, and *PAX9*).

NFKBIA (NFKB inhibitor alpha)—It has been shown to modulate *WNT*, *SHH*, and *LHX2IS* signaling at early stages of hair follicle development in mice. In particular, in the epidermis of mice lacking the transcription factor nuclear factor-kappa B activity, primary hair follicle pre-placode formation is initiated without progression to proper placodes (Tomann et al., 2016). The gene has been also detected as putatively under selection in a Chinese Merino sheep population (Liu et al., 2017).

SEC23A—It is one of the major components and markers of COPII vesicles from endoplasmic reticulum. It has been found associated with the sparse hair phenotype in humans (https://mseqdr.org/hpo_browser.php?8070) Moreover, it may contain causative mutations for an autosomal recessive disease known as cranio–lenticulo–sutural dysplasia, *alias* Boyadjiev–Jabs syndrome, in which patients have abnormal hair, among other cranio-facial abnormalities. Also, it has been shown to co-localize with the three proteins, transmembrane (Cdh23), scaffold (harmonin), and actin-based motor (Myo7a), whose defect is responsible of various types of the Usher syndrome, a multi-genic congenital disease characterized by perturbation of normal organization and growth of hair bundles within the inner ear (Blanco-Sánchez et al., 2014)

PAX9—It is a member of the paired box (PAX) family of transcription factors. Heterozygous mutations in *PAX9* have been associated in humans with non-syndromic tooth agenesis, non-syndromic, and familial oligodontia, with peg-shaped laterals and microdontia incisors. Often, these symptoms are associated with hair defects (Roberts and Chetty, 2018) as the same genes responsible for tooth development are involved in the growth and development of the other tissues derived from the ectoderm, including hair.

In general, biological evidence for the involvement in plausible Merino phenotypes was observed for the vast majority of coding genes in putatively selected regions detected either *via* "local" ancestry" or the F_{ST}-outlier" approach. The above result highlights the power of the multi-cohort approach adopted here. While we cannot exclude that false positive signals may have been retained in this study, still this represents so far the most complete genome-wide study of selection signatures for the Merino phenotype. The selection signatures reported here provide a comprehensive insight into the genetic basis underlining the Merino phenotype in sheep, which appeared here to be mainly represented by wool (and horn) traits. Targeted studies at both physiological and molecular levels will be needed to better understand the biological complexity behind these commercially relevant traits.

DATA AVAILABILITY STATEMENT

The whole SNP genotype dataset is available on the WIDDE database (http://widde.toulouse.inra.fr/widde/).

AUTHOR CONTRIBUTIONS

Conception of the work: SMe, EC. Data analysis: SMe, SMa, EC. Results interpretation: SMe, SMa, JL, EC. Drafting the article: SMe, EC. Critical revision of the article: SMe, SMa, MH, JL, EC. Final approval of the version to be published: SMe, SMa, MH, JL, EC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01025/full#supplementary-material

DATA SHEET 1 | Information on the considered breeds.

FIGURE S1 | Individual proportions of global admixture for the four considered datasets. Dataset composition: (A) Rambouillet, Chinese Merino, Tibetan; (B) Australian Merino, Sopravissana, Appenninica; (C) Gentile di Puglia, Spanish Merino, Appenninica; (D) Gentile di Puglia, Australian Merino, Appenninica. To

estimate admixture proportions of the four test breeds (A, Chinese Merino; B, Sopravissana; C, Spanish Merino; D, Australian Merino), each dataset was assumed to be arranged into two sub-populations. Color codes define the admixture proportions for each animal. Individual proportions of global admixture were averaged within breed to obtain a, the fraction of global admixture, adopted as parameter in the "local ancestry" analyses (see main text).

TABLE S1 | Sheep breeds considered throughout this study. N, number of genotyped animals.

TABLE S2 | Sheep breeds considered in the four different scenarios tested using the "local ancestry" approach. N, number of genotyped animals; α, fraction of global admixture.

TABLE S3 | Sheep breeds considered in the nine pair-wise comparisons tested using the "FST-outlier" approach.

TABLE S4 | Putatively selected regions identified in Australian Merino using the "local ancestry" approach. Merino genome-wide marker average ancestry (MAA) for this tested scenario was 0.87. For each region, the sheep chromosome (OAR), the name and the position, expressed in base pairs (bp), of the start and end SNPs (SNP ID), together with the MAA values representing the Merino fraction for the start and end SNPs, are provided. In bold and italics the region overlapping with ROHs detected within the Australian Merino breed. Highlighted in light yellow, the regions where at least one significant SNP in at least one pair-wise comparison of the "FST-outlier" method involving the corresponding Merino (or Merino-derived) breed was observed (see main text).

TABLE S5 | Putatively selected regions identified in Chinese Merino using the "local ancestry" approach. Merino genome-wide marker average ancestry (MAA) for this tested scenario was 0.91. For each region, the sheep chromosome (OAR), the name and the position, expressed in base pairs (bp), of the start and end SNPs (SNP ID), together with the MAA values representing the Merino fraction for the start and end SNPs, are provided. Highlighted in light yellow, the regions where at least one significant SNP in at least one pair-wise comparison of the "FST-outlier" method involving the corresponding Merino (or Merino-derived) breed was observed (see main text).

TABLE S6 | Putatively selected regions identified in Sopravissana using the "local ancestry" approach. Merino genome-wide marker average ancestry (MAA) for this tested scenario was 0.76. For each region, the sheep chromosome (OAR), the name and the position, expressed in base pairs (bp), of the start and end SNPs (SNP ID), together with the MAA values representing the Merino fraction for the start and end SNPs, are provided. Highlighted in light yellow, the regions where at least one significant SNP in at least one pair-wise comparison of the "FST-outlier" method involving the corresponding Merino (or Merino-derived) breed was observed (see main text).

TABLE S7 | Putatively selected regions identified in Spanish Merino using the "local ancestry" approach. Merino genome-wide marker average ancestry (MAA) for this tested scenario was 0.77. For each region, the sheep chromosome (OAR), the name and the position, expressed in base pairs (bp), of the start and end SNPs (SNP ID), together with the MAA values representing the Merino fraction for the start and end SNPs, are provided. In bold and italics the region overlapping with ROHs detected within the Spanish Merino breed. Highlighted in light yellow, the regions where at least one significant SNP in at least one pair-wise comparison of the "FST-outlier" method involving the corresponding Merino (or Merino-derived) breed was observed (see main text).

TABLE S8 | Results of the analyses performed using the "FST-outlier" approach for the nine pair-wise comparisons between Merino and non-Merino breeds. The sheep chromosome (OAR), the name (SNP ID) and the position, expressed in base pairs (bp), of the significant loci displaying qval < 0.05, and the corresponding FST values, are shown. In bold and italics, loci included within ROH islands detected in the same Merino (or Merino-derived) breed (see main text). Highlighted in light yellow, loci located within putatively selected regions identified, for the corresponding four considered breeds (Australian Merino, Spanish Merino, Sopravissana, Chinese Merino) using the "local ancestry" approach.

TABLE S9 | Gene content of the region on OAR17 detected as putatively selected by "local ancestry" analysis. N., sequential numbers. Gene ID, gene symbol.

TABLE S10 | Gene content of the region on OAR18 detected as putatively selected by "local ances

TABLE S11 | Quantitative Trait Loci (QTLs) known in sheep and mapped to regions detected in this study as putatively selected via "local ancestry" (A) and "FST-outlier" (B) analysis. QTL ID, QTL accession code at the NCBI Sheep Genome Data Viewer. OAR, sheep chromosome.

TABLE S12 | Gene content of the regions on OAR3 (A), OAR10 (B), OAR13 (C) and OAR19 (D) detected as putatively selected by "FST-outlier" analysis. N., sequential number. Gene ID, gene symbol.

TABLE S13 | Results of the gene network analysis. In italics, interactions between genes detected as putatively selected using the same approach (either LAMP or FST-outlier). In bold, interactions between genes detected as putatively selected using different approaches.

REFERENCES

Ahmed, Z. M., Yousaf, R., Lee, B. C., Khan, S. N., Lee, S., Lee, K., et al. (2011). Functional null mutations of MSRB3 encoding methionine sulfoxide reductase are associated with human deafness DFNB74. *Am. J. Hum. Genet.* 88, 19–29. doi: 10.1016/j.ajhg.2010.11.010

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

Allain, D., Elsen, J. M., Francois, D., Brunei, J. C., Weisbecker, J. L., Schibler, L., et al. (1998). A design aiming at detecting QTL controlling wool traits and other traits in the inra401 sheep line. *Proc. 6th World Congr. Genet. Appl. Livest. Prod.* 26, 434–436.

Allain, D., Schibler, L., Mura, L., Barillet, F., Sechi, T., Rupp, R., et al. (2006). QTL detection with DNA markers for wool traits in a sheep backcross Sarda × Lacaune resource population. *Proc. 8th World Congr. Genet. Appl. Livest. Prod.* 2006, 5–7.

Beh, K. J., Callaghan, M. J., Leish, Z., Hulme, D. J., Lenane, I., and Maddox, J. F. (2001). A genome scan for QTL affecting fleece and wool traits in Merino sheep. *Wool Tech. Sheep Breed.* 49, 88–97. doi: 10.1046/j.1365-2052.2002.00829.x

Bidinost, F., Roldan, D. L., Dodero, A. M., Cano, E. M., Taddeo, H. R., Mueller, J. P., et al. (2008). Wool quantitative trait loci in Merino sheep. *Small Rumin. Res.* 74, 113–118. doi: 10.1016/j.smallrumres.2007.04.005

Blanco-Sánchez, B., Clément, A., Fierro, J., Washbourne, P., and Westerfield, M. (2014). Complexes of Usher proteins preassemble at the endoplasmic reticulum and are required for trafficking and ER homeostasis. *Dis. Model. Mech.* 7, 547–559. doi: 10.1242/dmm.014068

Bolormaa, S., Swan, A. A., Brown, D. J., Hatcher, S., Moghaddar, N., van der Werf, J. H., et al. (2017). Multiple-trait QTL mapping and genomic prediction for wool traits in sheep. *Genet. Sel. Evol.* 49, 62. doi: 10.1186/s12711-017-0337-y

Carlson, T. L., Moini, M., Eckenrode, B. A., Allred, B. M., and Donfack, J. (2018). Protein extraction from human anagen head hairs 1-millimeter or less in total length. *BioTechniques* 64, 170–176. doi: 10.2144/btn-2018-2004

Chai, W., Zhou, H., Forrest, R. H. J., Gong, H., Hodge, S., and Hickford, J. G. H. (2017). Polymorphism of KRT83 and its association with selected wool traits in Merino-cross lambs. *Small Rumin. Res.* 155, 6–11. doi: 10.1016/j.smallrumres.2017.08.019

Chang, C.-Y. (2014) *Coordinating Stem Cell Behavior in the Hair Follicle*. Student Theses and Dissertations. 217.

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

Ciani, E., Lasagna, E., D'Andrea, M., Alloggio, I., Marroni, F., Ceccobelli, S., et al. (2015). Merino and Merino-derived sheep breeds: a genome-wide intercontinental study. *Genet. Sel. Evol.* 47, 64. doi: 10.1186/s12711-015-0139-z

Cong, J., Geng, W., He, B., Liu, J., Charlton, J., and Adler, P. N. (2001). The furry gene of Drosophila is important for maintaining the integrity of cellular extensions during morphogenesis. *Dev. Camb. Engl.* 128, 2793–2802.

Cornella, N., Tebaldi, T., Gasperini, L., Singh, J., Padgett, R. A., Rossi, A., et al. (2017). The hnRNP RALY regulates transcription and cell proliferation by modulating the expression of specific factors including the proliferation marker E2F1. *J. Biol. Chem.* 292, 19674–19692. doi: 10.1074/jbc.M117.795591

Covitz, P. A., Smith, L. S., and Long, S. R. (1998). Expressed sequence tags from a root-hair-enriched medicago truncatula cDNA library. *Plant Physiol.* 117, 1325–1332. doi: 10.1104/pp.117.4.1325

Debono Spiteri, C., Gillis, R. E., Roffet-Salque, M., Castells Navarro, L., Guilaine, J., Manen, C., et al. (2016). Regional asynchronicity in dairy production and processing in early farming communities of the northern Mediterranean. *Proc. Natl. Acad. Sci. U. S. A.* 113, 13594–13599. doi: 10.1073/pnas.1607810113

Demars, J., Cano, M., Drouilhet, L., Plisson-Petit, F., Bardou, P., Fabre, S., et al. (2017). Genome-wide identification of the mutation underlying fleece variation and discriminating ancestral hairy species from modern woolly sheep. *Mol. Biol. Evol.* 34, 1722–1729. doi: 10.1093/molbev/msx114

Demirci, S., Baştanlar, E. K., Dağtaş, N. D., Pişkin, E., Engin, A., Özer, F., et al. (2013). Mitochondrial DNA Diversity of Modern, Ancient and Wild Sheep (Ovis gmelinii anatolica) from Turkey: New Insights on the Evolutionary History of Sheep. *PLoS ONE* 8, e81952. doi: 10.1371/journal.pone.0081952

Dreger, D. L., Parker, H. G., Ostrander, E. A., and Schmutz, S. M. (2013). Identification of a mutation that is associated with the saddle tan and black-and-tan phenotypes in Basset Hounds and Pembroke Welsh Corgis. *J. Hered.* 104, 399–406. doi: 10.1093/jhered/est012

Duhl, D. M., Stevens, M. E., Vrieling, H., Saxon, P. J., Miller, M. W., Epstein, C. J., et al. (1994). Pleiotropic effects of the mouse lethal yellow (Ay) mutation explained by deletion of a maternally expressed gene and the simultaneous production of agouti fusion RNAs. *Dev. Camb. Engl.* 120, 1695–1708.

Dymova, M. A., Zadorozhny, A. V., Mishukova, O. V., Khrapov, E. A., Druzhkova, A. S., Trifonov, V. A., et al. (2017). Mitochondrial DNA analysis of ancient sheep from Altai. *Anim. Genet.* 48, 615–618. doi: 10.1111/age.12569

Ebrahimi, F., Gholizadeh, M., Rahimi-Mianji, G., and Farhadi, A. (2017). Detection of QTL for greasy fleece weight in sheep using a 50 K single nucleotide polymorphism chip. *Trop. Anim. Health Prod.* 49, 1657–1662. doi: 10.1007/s11250-017-1373-x

Ethier, J., Bánffy, E., Vuković, J., Leshtakov, K., Bacvarov, K., Roffet-Salque, M., et al. (2017). Earliest expansion of animal husbandry beyond the Mediterranean zone in the sixth millennium BC. *Sci. Rep.* 7, 7146. doi: 10.1038/s41598-017-07427-x

Fang, X., Lu, Q., Emoto, K., and Adler, P. N. (2010). The Drosophila Fry protein interacts with Trc and is highly mobile in vivo. *BMC Dev. Biol.* 10, 40. doi: 10.1186/1471-213X-10-40

Fariello, M.-I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., Consortium, I. S. G., et al. (2014). Selection signatures in worldwide sheep populations. *PLOS ONE* 9, e103813. doi: 10.1371/journal.pone.0103813

Foll, M., and Gaggiotti, O. (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

Furumura, M., and Ishikawa, H. (1996). Actin bundles in human hair follicles as revealed by confocal laser microscopy. *Cell Tissue Res.* 283, 425–434. doi: 10.1007/s004410050553

García-Gámez, E., Reverter, A., Whan, V., McWilliam, S. M., Arranz, J. J.International Sheep Genomics Consortium, et al. (2011). Using regulatory and epistatic networks to extend the findings of a genome scan: identifying the gene drivers of pigmentation in merino sheep. *PloS One* 6, e21158. doi: 10.1371/journal.pone.0021158

Gong, H., Zhou, H., Bai, L., Li, W., Li, S., Wang, J., et al. (2019). Associations between variation in the ovine high glycine-tyrosine keratin-associated protein gene KRTAP20-1 and wool traits. *J. Anim. Sci.* 97, 587–595. doi: 10.1093/jas/sky465

Guild, G. M., Connelly, P. S., Ruggiero, L., Vranich, K. A., and Tilney, L. G. (2005). Actin filament bundles in Drosophila wing hairs: hairs and bristles use different strategies for assembly. *Mol. Biol. Cell* 16, 3620–3631. doi: 10.1091/mbc.e05-03-0185

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

Gutiérrez-Gil, B., Esteban-Blanco, C., Wiener, P., Chitneedi, P. K., Suarez-Vega, A., and Arranz, J.-J. (2017). High-resolution analysis of selection sweeps identified between fine-wool Merino and coarse-wool Churra sheep breeds. *Genet. Sel. Evol.* 49, 81. doi: 10.1186/s12711-017-0354-x

Harmon, C. S., and Nevins, T. D. (1997). Evidence that activation of protein kinase A inhibits human hair follicle growth and hair fibre production in organ culture and DNA synthesis in human and mouse hair follicle organ culture. *Br. J. Dermatol.* 136 (6), 853–858. doi: 10.1111/j.1365-2133.1997.tb03924.x

He, Y., Fang, X., Emoto, K., Jan, Y.-N., and Adler, P. N. (2005). The tricornered Ser/Thr protein kinase is regulated by phosphorylation and interacts with furry during Drosophila wing hair development. *Mol. Biol. Cell* 16, 689–700. doi: 10.1091/mbc.e04-09-0828

Henry, H., Doods, K., Wuliji, T., Jenkis, Z., Beattie, A., and Montgomery, G. (1998). A genome screen for QTL for wool traits in a Merino x Romney backcross flock (Reprinted). *Wool Technol. Sheep Breed.* 46, 213–217.

Hung, R.-J., Spaeth, C. S., Yesilyurt, H. G., and Terman, J. R. (2013). SelR reverses Mical-mediated oxidation of actin to regulate F-actin dynamics. *Nat. Cell Biol.* 15, 1445–1454. doi: 10.1038/ncb2871

Ivanova, M., De Cupere, B., Ethier, J., and Marinova, E. (2018a). Correction: Pioneer farming in southeast Europe during the early sixth millennium BC: Climate-related adaptations in the exploitation of plants and animals. *PloS One* 13, e0202668. doi: 10.1371/journal.pone.0202668

Ivanova, M., De Cupere, B., Ethier, J., and Marinova, E. (2018b). Pioneer farming in southeast Europe during the early sixth millennium BC: Climate-related adaptations in the exploitation of plants and animals. *PloS One* 13, e0197225. doi: 10.1371/journal.pone.0197225

Jacobs, L. C., Hamer, M. A., Gunn, D. A., Deelen, J., Lall, J. S., van Heemst, D., et al. (2015). A genome-wide association study identifies the skin color genes IRF4, MC1R, ASIP, and BNC2 influencing facial pigmented spots. *J. Invest. Dermatol.* 135, 1735–1742. doi: 10.1038/jid.2015.62

Johnston, S. E., McEwan, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., et al. (2011). Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Mol. Ecol.* 20, 2555–2566. doi: 10.1111/j.1365-294X.2011.05076.x

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Neto, L. R. P., Cristobal, M. S., 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

Kim, S., Wong, P., and Coulombe, P. A. (2006). A keratin cytoskeletal protein regulates protein synthesis and epithelial cell growth. *Nature* 441, 362–365. doi: 10.1038/nature04659

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 42, D966–D974. doi: 10.1093/nar/gkt1026

Lee, Y. J., Rice, R. H., and Lee, Y. M. (2006). Proteome analysis of human hair shaft: from protein identification to posttranslational modification. *Mol. Cell. Proteomics* 5, 789–800. doi: 10.1074/mcp.M500278-MCP200

Li, J., Zhao, X., Xin, Q., Shan, S., Jiang, B., Jin, Y., et al. (2015). Whole-exome sequencing identifies a variant in TMEM132E causing autosomal-recessive non syndromic hearing loss DFNB99. *Hum. Mutat.* 36, 98–105. doi: 10.1002/humu.22712

Li, S., Zhou, H., Gong, H., Zhao, F., Hu, J., Luo, Y., et al. (2017a). Identification of the Ovine Keratin-Associated Protein 26-1 Gene and Its Association with Variation in Wool Traits. *Genes* 8. doi: 10.3390/genes8090225

Li, S., Zhou, H., Gong, H., Zhao, F., Wang, J., Liu, X., et al. (2017b). Identification of the Ovine Keratin-Associated Protein 22-1 (KAP22-1) Gene and Its Effect on Wool Traits. *Genes* 8. doi: 10.3390/genes8010027

Ling, Y. H., Xiang, H., Zhang, G., Ding, J. P., Zhang, Z. J., Zhang, Y. H., et al. (2014). Identification of complete linkage disequilibrium in the DSG4 gene and its association with wool length and crimp in Chinese indigenous sheep. *Genet. Mol. Res.* 13, 5617–5625. doi: 10.4238/2014.July.25.17

Liu, G., Liu, R., Li, Q., Tang, X., Yu, M., Li, X., et al. (2013). Identification of microRNAs in wool follicles during anagen, catagen, and telogen phases in Tibetan sheep. *PloS One* 8, e77801. doi: 10.1371/journal.pone.0077801

Liu, G., Liu, R., Tang, X., Cao, J., Zhao, S., and Yu, M. (2015). Expression profiling reveals genes involved in the regulation of wool follicle bulb regression and regeneration in sheep. *Int. J. Mol. Sci.* 16, 9152–9166. doi: 10.3390/ijms16059152

Liu, S., He, S., Chen, L., Li, W., Di, J., and Liu, M. (2017). Estimates of linkage disequilibrium and effective population sizes in Chinese Merino (Xinjiang type) sheep by genome-wide SNPs. *Genes Genomics* 39, 733–745. doi: 10.1007/s13258-017-0539-2

Ma, G.-W., Chu, Y.-K., Zhang, W.-J., Qin, F.-Y., Xu, S.-S., Yang, H., et al. (2017). Polymorphisms of FST gene and their association with wool quality traits in Chinese Merino sheep. *PloS One* 12, e0174868. doi: 10.1371/journal.pone.0174868

Manunza, A., Cardoso, T. F., Noce, A., Martínez, A., Pons, A., Bermejo, L. A., et al. (2016). Population structure of eleven Spanish ovine breeds and detection of selective sweeps with BayeScan and hapFLK. *Sci. Rep.* 6, 27296. doi: 10.1038/srep27296

Mastrangelo, S., Sardina, M. T., Tolone, M., Di Gerlando, R., Sutera, A. M., Fontanesi, L., et al. (2018). Genome-wide identification of runs of homozygosity islands and associated genes in local dairy cattle breeds. *Animal* 12, 2480–2488. doi: 10.1017/S1751731118000629

Mastrangelo, S., Bahbahani, H., Moioli, B., Ahbara, A., Abri, M. A., Almathen, F., et al. (2019). Novel and known signals of selection for fat deposition in domestic sheep breeds from Africa and Eurasia. *PLoS One* 14, e0209632. doi: 10.1371/journal.pone.0209632

Mastrangelo, S., Tolone, M., Sardina, M. T., Sottile, G., Sutera, A. M., Di Gerlando, R., et al. (2017). Genome-wide scan for runs of homozygosity identifies potential candidate genes associated with local adaptation in Valle del Belice sheep. *Genet. Sel. Evol.* 49, 84. doi: 10.1186/s12711-017-0360-z

Meadows, J. R. S., Hiendleder, S., and Kijas, J. W. (2011). Haplogroup relationships between domestic and wild sheep resolved using a mitogenome panel. *Heredity* 106, 700–706. doi: 10.1038/hdy.2010.122

Metzger, J., Karwath, M., Tonda, R., Beltran, S., Águeda, L., Gut, M., et al. (2015). Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics* 16, 764. doi: 10.1186/s12864-015-1977-3

Michaud, E. J., Bultman, S. J., Stubbs, L. J., and Woychik, R. P. (1993). The embryonic lethality of homozygous lethal yellow mice (Ay/Ay) is associated with the disruption of a novel RNA-binding protein. *Genes Dev.* 7, 1203–1213. doi: 10.1101/gad.7.7a.1203

Moioli, B., Pilla, F., and Ciani, E. (2015). Signatures of selection identify loci associated with fat tail in sheep. *J. Anim. Sci.* 93, 4660–4669. doi: 10.2527/jas.2015-9389

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology* 9, S4. doi: 10.1186/gb-2008-9-s1-s4

Mu, F., Rong, E., Jing, Y., Yang, H., Ma, G., Yan, X., et al. (2017). Structural characterization and association of ovine Dickkopf-1 gene with wool production and quality traits in Chinese Merino. *Genes* 8. doi: 10.3390/genes8120400

Nasr, Z., Dow, L. E., Paquet, M., Chu, J., Ravindar, K., Somaiah, R., et al. (2013). Suppression of eukaryotic initiation factor 4E prevents chemotherapy-induced alopecia. *BMC Pharmacol. Toxicol.* 14, 58. doi: 10.1186/2050-6511-14-58

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

Pan, Z., Li, S., Liu, Q., Wang, Z., Zhou, Z., Di, R., et al. (2018). Whole-genome sequences of 89 Chinese sheep suggest role of RXFP2 in the development of unique horn phenotype as response to semi-feralization. *GigaScience* 7. doi: 10.1093/gigascience/giy019

Parsons, Y. M., Cooper, D. W., and Piper, L. R. (1994). Evidence of linkage between high-glycine-tyrosine keratin gene loci and wool fibre diameter in a Merino half-sib family. *Anim. Genet.* 25, 105–108. doi: 10.1111/j.1365-2052.1994.tb00088.x

Phuaa, S. H., Scobieb, D. R., O'Connellb, D., Henrya, H., Doddsa, K. G., and Brauninga, R. (2015). Preliminary linkage studies in sheep of keratin and keratin-associated protein genes with fleece weight, wool fibre diameter and fibre curvature. *Proc. N. Z. Soc. Anim. Prod.* 75, 101–105.

Ponz, R., Moreno, C., Allain, D., Elsen, J. M., Lantier, F., Lantier, I., et al. (2001). Assessment of genetic variation explained by markers for wool traits in sheep via a segment mapping approach. *Mamm. Genome* 12, 569–572. doi: 10.1007/s003350030007

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Purfield, D. C., McParland, S., Wall, E., and Berry, D. P. (2017). The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. *PloS One* 12, e0176780. doi: 10.1371/journal.pone.0176780

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T. M., Fries, R., et al. (2014). Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.* 10, e1004148. doi: 10.1371/journal.pgen.1004148

Roberts, M. R., Asgari, M. M., and Toland, A. E. (2019). Genome-wide association studies and polygenic risk scores for skin cancer: clinically useful yet? *Br. J. Dermatol.* doi: 10.1111/bjd.17917

Roberts, T. S., and Chetty, M. (2018). Hypohidrotic Ectodermal Dysplasia: Genetic aspects and clinical implications of hypodontia. *South Afr. Dent. J.* 73, 253–256.

Rochus, C. M., Tortereau, F., Plisson-Petit, F., Restoux, G., Moreno-Romieux, C., Tosser-Klopp, G., et al. (2018). Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. *BMC Genomics* 19, 71. doi: 10.1186/s12864-018-4447-x

Rogers, G. R., Hickford, J. G. H., and Bickerstaffe, R. (1994). A potential QTL for wool strength located on ovine chromosome 11. *Proc. 5th World Congr. Genet. Appl. Livest. Prod.* 21, 291–294.

Roldan, D. L., Dodero, A. M., Bidinost, F., Taddeo, H. R., Allain, D., Poli, M. A., et al. (2010). Merino sheep: a further look at quantitative trait loci for wool production. *Animal* 4, 1330–1340. doi: 10.1017/S1751731110000315

Rong, E. G., Yang, H., Zhang, Z. W., Wang, Z. P., Yan, X. H., Li, H., et al. (2015). Association of methionine synthase gene polymorphisms with wool production and quality traits in Chinese Merino population. *J. Anim. Sci.* 93, 4601–4609. doi: 10.2527/jas.2015-8963

Rossi, A., Moro, A., Tebaldi, T., Cornella, N., Gasperini, L., Lunelli, L., et al. (2017). Identification and dynamic changes of RNAs isolated from RALY-containing ribonucleoprotein complexes. *Nucleic Acids Res.* 45, 6775–6792. doi: 10.1093/nar/gkx235

Ryder, M. (1981). A survey of European primitive breeds of sheep. *Ann. Genet. Sel. Anim.* 13, 381–418. doi: 10.1186/1297-9686-13-4-381

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* 82, 290–303. doi: 10.1016/j.ajhg.2007.09.022

Sasikumar, A. N., Perez, W. B., and Kinzy, T. G. (2012). The many roles of the eukaryotic elongation factor 1 complex. *Wiley Interdiscip. Rev. RNA* 3, 543–555. doi: 10.1002/wrna.1118

Seroussi, E., Rosov, A., Shirak, A., Lam, A., and Gootwine, E. (2017). Unveiling genomic regions that underlie differences between Afec-Assaf sheep and its parental Awassi breed. *Genet. Sel. Evol.* 49, 19. doi: 10.1186/s12711-017-0296-3

Shen, X., Liu, F., Wang, Y., Wang, H., Ma, J., Xia, W., et al. (2015). Down-regulation of msrb3 and destruction of normal auditory system development through hair cell apoptosis in zebrafish. *Int. J. Dev. Biol.* 59, 195–203. doi: 10.1387/ijdb.140200md

Shriner, D. (2013). Overview of admixture mapping. *Curr. Protoc. Hum. Genet.* 76, 1.23.1–1.23.8. doi: 10.1002/0471142905.hg0123s76

Singh, S., Kumar, S., Kolte, A. P., and Kumar, S. (2013). Extensive variation and sub-structuring in lineage A mtDNA in Indian sheep: genetic evidence for domestication of sheep in India. *PloS One* 8, e77858. doi: 10.1371/journal.pone.0077858

Sulayman, A., Tursun, M., Sulaiman, Y., Huang, X., Tian, K., Tian, Y., et al. (2018). Association analysis of polymovwwrphisms in six keratin genes with wool traits in sheep. *Asian-Australas. J. Anim. Sci.* 31, 775–783. doi: 10.5713/ajas.17.0349

Talenti, A., Bertolini, F., Pagnacco, G., Pilla, F., Ajmone-Marsan, P., Rothschild, M. F., et al. (2017a). The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* 28, 114–128. doi: 10.1007/s00335-017-9678-7

Talenti, A., Bertolini, F., Pagnacco, G., Pilla, F., Ajmone-Marsan, P., Rothschild, M. F., et al. (2017b). Erratum to: The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions. *Mamm. Genome* 28, 129. doi: 10.1007/s00335-017-9685-8

Taungjaruwinai, W. M., Bhawan, J., Keady, M., and Thiele, J. J. (2009). Differential expression of the antioxidant repair enzyme methionine sulfoxide reductase (MSRA and MSRB) in human skin. *Am. J. Dermatopathol.* 31, 427–431. doi: 10.1097/DAD.0b013e3181882c21

Tomann, P., Paus, R., Millar, S. E., Scheidereit, C., and Schmidt-Ullrich, R. (2016). Lhx2 is a direct NF-κB target gene that promotes primary hair follicle placode down-growth. *Development* 143, 1512–1522. doi: 10.1242/dev.130898

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S., et al. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.* 7, e1002316. doi: 10.1371/journal.pgen.1002316

Vermorken, A. J., Weterings, P. J., Kibbelaar, M. A., Lenstra, J. A., and Bloemendal, H. (1981). Isolation and characterization of actin from human hair follicles. *FEBS Lett.* 127, 105–108. doi: 10.1016/0014-5793(81)80352-3

Wang, H., Zhang, L., Cao, J., Wu, M., Ma, X., Liu, Z., et al. (2015). Genome-wide specific selection in three domestic sheep breeds. *PloS One* 10, e0128688. doi: 10.1371/journal.pone.0128688

Wang, Z., Zhang, H., Yang, H., Wang, S., Rong, E., Pei, W., et al. (2014). Genome-wide association study for wool production traits in a Chinese Merino sheep population. *PloS One* 9, e107101. doi: 10.1371/journal.pone.0107101

Wei, C., Wang, H., Liu, G., Wu, M., Cao, J., Liu, Z., et al. (2015). Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genomics* 16, 194. doi: 10.1186/s12864-015-1384-9

Wilkinson, S., Lu, Z. H., Megens, H.-J., Archibald, A. L., Haley, C., Jackson, I. J., et al. (2013). Signatures of diversifying selection in European pig breeds. *PLoS Genet.* 9, e1003453. doi: 10.1371/journal.pgen.1003453

Yang, J., Li, W.-R., Lv, F.-H., He, S.-G., Tian, S.-L., Peng, W.-F., et al. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33, 2576–2592. doi: 10.1093/molbev/msw129

Yuan, Z., Liu, E., Liu, Z., Kijas, J. W., Zhu, C., Hu, S., et al. (2017). Selection signature analysis reveals genes associated with tail type in Chinese indigenous sheep. *Anim. Genet.* 48, 55–66. doi: 10.1111/age.12477

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci.* 105, 11597–11604. doi: 10.1073/pnas.0801317105

Zhai, M., Xie, Y., Yang, M., Mu, J., and Zhao, Z. (2019). Investigation of the relationships between wool quality and microsatellite in hybrids of Australian Merino and Chinese Merino. *Kafkas Univ. Vet. Fak. Derg.* 25 (2), 163–170. doi: 10.9775/kvfd.2018.2020488

Zhang, L., Mousel, M. R., Wu, X., Michal, J. J., Zhou, X., Ding, B., et al. (2013). Genome-wide genetic diversity and differentially selected regions among Suffolk, Rambouillet, Columbia, Polypay, and Targhee sheep. *PloS One* 8, e65942. doi: 10.1371/journal.pone.0065942

Zhou, C., Zang, D., Jin, Y., Wu, H., Liu, Z., Du, J., et al. (2011). Mutation in ribosomal protein L21 underlies hereditary hypotrichosis simplex. *Hum. Mutat.* 32, 710–714. doi: 10.1002/humu.21503

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.

The handling editor declared a past co-authorship with one of the authors SMa.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 November 2019
doi: 10.3389/fgene.2019.01170

Mapping Recombination Rate on the Autosomal Chromosomes Based on the Persistency of Linkage Disequilibrium Phase Among Autochthonous Beef Cattle Populations in Spain

*Elena Flavia Mouresan[1], Aldemar González-Rodríguez[1], Jhon Jacobo Cañas-Álvarez[2], Sebastián Munilla[1,3], Juan Altarriba[1], Clara Díaz[4], Jesús A. Baró[5], Antonio Molina[6], Pascual Lopez-Buesa[1], Jesús Piedrafita[2] and Luis Varona[1]**

[1] Departamento de Anatomía, Embriología y Genética Animal, Universidad de Zaragoza, Zaragoza, Spain, [2] Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Barcelona, Spain, [3] Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, CONICET, Buenos Aires, Argentina, [4] Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, Spain, [5] Instituto Agroalimentario de Aragón (IA2), Zaragoza, Spain, [6] Departamento de Ciencias Agroforestales, Universidad de Valladolid, Valladolid, Spain, [7] Departamento de Genética, Universidad de Córdoba, Córdoba, Spain, [8] Departamento de Producción Animal y Ciencia de los Alimentos, Universidad de Zaragoza, Zaragoza, Spain

In organisms with sexual reproduction, genetic diversity, and genome evolution are governed by meiotic recombination caused by crossing-over, which is known to vary within the genome. In this study, we propose a simple method to estimate the recombination rate that makes use of the persistency of linkage disequilibrium (LD) phase among closely related populations. The biological material comprised 171 triplets (sire/dam/offspring) from seven populations of autochthonous beef cattle in Spain (Asturiana de los Valles, Avileña-Negra Ibérica, Bruna dels Pirineus, Morucha, Pirenaica, Retinta, and Rubia Gallega), which were genotyped for 777,962 SNPs with the BovineHD BeadChip. After standard quality filtering, we reconstructed the haplotype phases in the parental individuals and calculated the LD by the correlation *-r-* between each pair of markers that had a genetic distance < 1 Mb. Subsequently, these correlations were used to calculate the persistency of LD phase between each pair of populations along the autosomal genome. Therefore, the distribution of the recombination rate along the genome can be inferred since the effect of the number of generations of divergence should be equivalent throughout the genome. In our study, the recombination rate was highest in the largest chromosomes and at the distal portion of the chromosomes. In addition, the persistency of LD phase was highly heterogeneous throughout the genome, with a ratio of 25.4 times between the estimates of the recombination rates from the genomic regions that had the highest (BTA18-7.1 Mb) and the lowest (BTA12-42.4 Mb) estimates. Finally, an overrepresentation enrichment analysis (ORA) showed

OPEN ACCESS

Edited by:
Francesca Bertolini, Technical University of Denmark, Denmark

Reviewed by:
Alessandro Bagnato, University of Milan, Italy
Diercles Francisco Cardoso, São Paulo State University, Brazil
Andressa Oliveira De Lima, Federal University of São Carlos, Brazil
Christos Palaiokostas, Swedish University of Agricultural Sciences, Sweden

***Correspondence:**
Luis Varona
lvarona@unizar.es

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *16 November 2018*
Accepted: *23 October 2019*
Published: *20 November 2019*

Citation:
Mouresan EF, González-Rodríguez A, Cañas-Álvarez JJ, Munilla S, Altarriba J, Díaz C, Baró JA, Molina A, Lopez-Buesa P, Piedrafita J and Varona L (2019) Mapping Recombination Rate on the Autosomal Chromosomes Based on the Persistency of Linkage Disequilibrium Phase Among Autochthonous Beef Cattle Populations in Spain. Front. Genet. 10:1170. doi: 10.3389/fgene.2019.01170

differences in the enriched gene ontology (GO) terms between the genes located in the genomic regions with estimates of the recombination rate over (or below) the 95th (or 5th) percentile throughout the autosomal genome.

Keywords: recombination rate, linkage disequilibrium, beef cattle, multiple populations, gene ontology

INTRODUCTION

Recombination caused by crossing-over during meiosis play a crucial role in the genetic diversity and the genome evolution of organisms with sexual reproduction (Arnheim et al., 2007). It creates new genetic variation by generating novel combinations of grand-paternal and grand-maternal genetic information, and it helps to remove deleterious mutations that might otherwise accumulate (Tiemann-Boege et al., 2017).

In most studies of genome-wide association or genomic selection, the distribution of crossing-over events had been considered uniform, although there is strong evidence that recombination rate is heterogeneous along the genome (Myers et al., 2005; Stapley et al., 2017). In general, recombination is higher in the regions of the telomeres and smaller near the centromere (Coop and Przeworski, 2007; Ma et al., 2015). Due to recombination, the genome is organized into haplotype blocks of varying lengths, as described in humans (Gabriel et al., 2002) and other species as rat and mouse (Guryev et al., 2006) and cattle (Mokry et al., 2014). The reason of this structure is the presence of small genomic regions that have a higher rate of recombination, known as recombination hotspots (Paigen and Petkov, 2010).

In addition, patterns of the recombination rate throughout the genome vary among species, populations, or even within individuals in different environments (Stapley et al., 2017). The evolution of the distribution of the recombination rate along the genome is an active research field (Dapper and Payseur, 2017). In general, it differs according to the genomic scale in which the recombination rate is measured (Smukowski and Noor, 2011). In a very fine scale (few kb), a rapid divergence of the recombination rate between mammal populations is observed (Auton et al., 2012; Stevison et al., 2016), whereas greater correlations are observed between closely related populations when they are calculated through larger chromosomal segments (Smukowski and Noor, 2011; Shen et al., 2018).

Traditionally, the distribution of the crossing-overs or recombination events within the genome has been studied by counting the number of chiasmata during meiosis (Hulten et al., 1982) or from linkage maps created from a limited number of genetic markers or phenotypes (Sturtevant, 1913). In recent years, high-throughput sequencing and genotyping technologies have provided a valuable new tool for measuring recombination rates with two main group of methods. First, estimates of recombination rates are based on observations of recombination events in large pedigrees between pairs of parent-offspring genotypes (Kong et al., 2010; Ma et al., 2015; Shen et al., 2018) or in sperm typing (Sarbajna et al., 2012) and require genotypic information from a large number of families or sperm cells. Second, other methods are based on the identification of local patterns in linkage disequilibrium (LD) with coalescent methods (McVean et al., 2002; Li and Stephens, 2003; Wall and Stevison, 2016), which estimate the background recombination rate $\rho_w=4N_wc_w$ −, where N_w and c_w are the indistinguishable effective population size and recombination rate for a specific window of the genome, respectively. The main limitation of the last approach is that the effective population size can vary dramatically over time. In fact, the decay of LD has been used to estimate past population history in humans (Hayes et al., 2003; Tenesa et al., 2007; Park, 2012) and livestock populations (Hayes et al., 2003; De Roos et al., 2008; Xu et al., 2019).

In addition, stratification of the population can severely distort the estimates of recombination rates because subdivisions of the population have a strong effect on LD estimates (Hinrichs et al., 2009). After reproductive isolation, the structure of LD tends to differ between subpopulations and the similarity (or persistency) of those LD patterns depends on the number of generations of divergence and the recombination rate plus other evolutionary events such as admixture or variations on the effective size of the populations (Hill and Robertson, 1968). For this reason, if genotypic information is available for closely related populations, measures of genome-wide persistency of LD phase among populations throughout the genome can be used to infer the distribution of recombination rate. The rationale of this approach is that genetic drift, admixture, or variations of the effective size should affect the entire genome with similar intensity and the heterogeneity of the persistency of LD phase is linked to variations on the recombination rate.

Therefore, the objective of this study was to develop a procedure to infer the distribution of the recombination rate from the persistency of LD phase among closely related populations and to apply it to genotypic data from seven beef cattle populations in Spain.

MATERIALS AND METHODS

The genomic data comprised the *BovineHD Genotyping Beadchip (777,962 SNPs, Illumina)* genotypes from 171 non-related triplets of sire, dam, and one offspring from seven breeds, being 25 *Asturiana de los Valles* (AV), 24 *Avileña - Negra Ibérica* (ANI), 25 *Bruna dels Pirineus* (BP), 25 *Morucha* (Mo), 24 *Pirenaica* (Pi), 24 *Retinta* (Re), and 24 *Rubia Gallega* (RG) triplets. This dataset has been used to analyze genetic differentiation (Cañas-Álvarez et al., 2015; Cañas-Álvarez et al., 2016; González-Rodríguez et al., 2017), signatures of selection (González-Rodríguez et al., 2016),

and haplotype diversity (Mouresan et al., 2017). These breeds represent 72% of the total census of local beef breeds in Spain (Ministerio de Medio Ambiente y Medio Rural y Marino, 2010) and their production systems are extensive or semi-extensive. The populations are reared in mountainous regions near the Pyrenees (*Pirenaica* and *Bruna dels Pirineus*) in the humid regions in northwestern Spain (*Rubia Gallega* and *Asturiana de los Valles*) or in pastures in semi-arid zones of the west and southwest of Spain (*Retinta*, *Avileña Negra-Ibérica*, and *Morucha*). The breeds differ in production and carcass traits (Piedrafita et al., 2003) and in their meat quality (Gil et al., 2001).

The triplets were sampled by the breeders associations with the aim of capturing most of the genetic variability of each population. We used an *ad-hoc* procedure that started with one triplet and incorporated the new ones by minimizing the total coancestry between them. The SNP filtering process included the following: 1) Mendelian error < 0.05, 2) SNP and individual call rates > 95%, and 3) Minor allele frequency (MAF) > 0.05 in pairs of populations. Only the SNPs that were located on autosomal chromosomes were retained. The filtering process yielded approximately 550,000 segregating markers for each pair of populations (see **Table 1**).

The genomic information of the triplets was used to reconstruct the parental haplotypes with the TRIO option of the *BEAGLE* software (Browning and Browning, 2007), which were used to calculate the LD in each population and between each pair of markers (i.e. with alleles A and a, and B and b, respectively) that had a genomic distance < 1 Mb. LD was estimated as a correlation –*r*-, as follows:

$$r = \frac{D}{\sqrt{p_A p_a p_B p_b}}$$

where $D = P_{AB}\, P_{ab}\, P_{Ab} P_{aB}$ (Falconer and Mackay, 1996), P_{AB} , P_{ab}, P_{Ab} and P_{ab} were the haplotype frequencies, and P_A, P_a, P_B, and P_b were the allelic frequencies.

To estimate the persistency of LD phase between pairs of populations, the Pearson correlations between LD estimates in each bin of 20 kb (0–20, 20–40, 40–60, 960–980, 980–1,000) within a 1 Mb window were calculated for each pair of populations. We obtained 50 correlation estimates (one per bin) between the LD estimates of each breed pair per window.

Under the assumption of constant variance of *r* (or effective population size) in both populations, the expectation of the correlation between LD estimates from a pair of SNP markers is e^{-2cT} (Hill and Robertson, 1968; De Roos et al., 2008), where *c* is the recombination rate between the markers and T is the number of generations of divergence between populations. Initially, it was assumed that the recombination rate was 1.25 cM per Mb (Arias et al., 2009). The regression of the natural logarithm of the correlations on the genomic distance was calculated and the slope was equated to -2cT to estimate T between each pair of populations.

Once the numbers of generations of divergence (T) were estimated from all available SNP markers, they were assumed as known and replaced by their estimates. Subsequently, the same expression (-2cT) was used to estimate *c*, although the analysis was restricted only to the SNP markers within 1 Mb, which were centered every 0.1 Mb along the autosomal genome in sliding windows. Therefore, 25,098 estimates of c were calculated for each of the 21 population pairs.

Afterwards, the presence of a common pattern for the distribution of the recombination rate was checked calculating the correlation between the estimates from all population pairs. Next, the estimates of *c* were averaged by chromosome, relative position within the chromosome, and location within the genome to infer the distribution of the recombination rate (or the persistency of LD phase) throughout the bovine autosomal genome. The degree of inequality of the recombination rate along the autosomal genome was measured with the Gini index –G- (Ceriani and Verme, 2012) as:

$$G = \frac{\sum_{i=1}^{ng}\sum_{j=1}^{ng}\left|c_i - c_j\right|}{2ng^2\overline{c}}$$

Where c_i and c_j were the average estimates of the recombination rate for the 21 population pairs ad for the *ith* and *jth* genomic regions, *ng* was the total number of genomic regions (25,098) defined along the autosomal genome, and $\overline{c}$ was the average of the 25,098 estimates of the recombination rate.

Finally, the 5th and 95th percentiles of the average recombination rates between pairs of populations in 0.1 Mb steps throughout the autosomal genome were calculated. Thereinafter, we prospected genes mapped within genomic regions falling out the 5–95th the percentile using the *Biomart* tool of *Ensembl* (Flicek et al., 2013) (www.ensembl.org). Further, we performed an Overrepresentation Enrichment Analysis (ORA) to determine if the overrepresentation of gene ontology (GO) terms differed between the two tails of the empirical distribution of recombination rates. We used the WEB-based Gene SeT AnaLysis Toolkit (www.webgestalt.org) using the *Homo sapiens* and *Bos taurus* annotation databases and with the complete genome as the reference set.

TABLE 1 | Number of co-segregating SNP markers between all possible pairs of seven beef cattle populations in Spain [Asturiana de los Valles (AV), Avileña - Negra Ibérica (ANI), Bruna dels Pirineus (BP), Morucha (Mo), Pirenaica (P), Retinta (Re), Rubia Gallega (RG)].

Pairs of populations	N° SNP markers	Pairs of populations	N° SNP markers
AV-ANI	555,373	**BP-Mo**	543,305
AV-BP	557,588	**BP-Pi**	534,336
AV-Mo	555,769	**BP-Re**	535,997
AV-Pi	540,390	**BP-RG**	544,350
AV-Re	547,893	**Mo-Pi**	529,281
AV-RG	553,868	**Mo-Re**	541,225
ANI-BP	538,327	**Mo-RG**	542,682
ANI-Mo	545,324	**Pi-Re**	522,670
ANI-Pi	524,630	**Pi-RG**	529,577
ANI-Re	536,595	**Re-RG**	535,677
ANI-RG	537,882		

RESULTS AND DISCUSSION

We have used the persistency of LD phase among seven closely related populations (Cañas-Álvarez et al., 2015) to infer the landscape of the recombination events in a sequential procedure that involved several steps. As in other studies in livestock populations (Brito et al., 2015; Biegelmeyer et al., 2016; Brito et al., 2017; Grossi et al., 2017), the similitude of LD among populations was very high between adjacent markers and decreased rapidly with genomic distance. In the first step, we used all the SNP markers to estimate the persistency of LD phase for each pair of populations as the slope of the regression analysis between the natural logarithm of the correlations between the *r* measures of LD in bins of 20 kb on the genomic distance. Theoretically, this relationship should be linear (Hill and Robertson, 1968; De Roos et al., 2008). However, the results varied according to the genomic distance evaluated (1 Mb, 500 Kb, 250 Kb, or 100 Kb). To illustrate this phenomenon, the results of the regression analysis between Re and RG are shown in **Figure 1**. The regression analysis of the data within a range of 250 Kb had the highest adjusted R^2 (0.999) and the linear relationship between persistency of LD phase and genomic distance was evident only in those first 250 Kb. The results for the other population pairs were similar (**Supplementary Information**, **Figures S1** to **S20**) and all of them had the highest adjusted R^2 (> 0.998) with the first 250 Kb. In all population pairs, persistency of LD phase decayed rapidly over short distances, but in larger genomic distances remained > 0, as observed by De Roos et al. (2008). This is probably due to the fact that the populations were not totally divergent or due to the presence of some migration between them (De Roos et al., 2008) and consistent with a decrease in effective population size in cattle (Hayes et al., 2003).

In a second step, we restricted the analysis to the LD within 250 Kb and the slope of the regression analyses were equated to -2*c*T, with *c* set to 1.25 cM per Mb (Arias et al., 2009). We obtained 21 estimates of the number of generations of divergence (T) between populations (**Table 2**), that ranged from 132.3 (AV-BP) to 281.9 (Pi-Re). Estimates were in concordance with the results obtained by Cañas-Álvarez et al. (2016) for the same populations and dataset, and by De Roos et al. (2008) between two dairy populations (Holstein-Friesian and Jersey). However, the divergence times found in our study were lower than the observed between a dairy (Holstein) and beef (Angus) populations (De Roos et al., 2008).

The estimates of T assumed that the variance of the LD (r) remained constant in each population and this is probably far from the truth. The effective size of the populations has decreased in the last generations (Cañas-Alvarez et al., 2016) and, thus, the estimates of T are probably overestimated. Nevertheless, the bias caused by the variations in the effective size should be similar throughout the autosomal genome and, therefore, local variations in the persistency of LD phase are informative for the inference of the recombination landscape along the autosomal genome.

After this preliminary step, we used the same expression (-2*c*T) to infer the distribution of the recombination rate *c*. Now, the numbers of generations of divergence (T) were assumed to be known and they were replaced by their estimates with all the SNP markers. In this case, we equated the slope of the regression of the

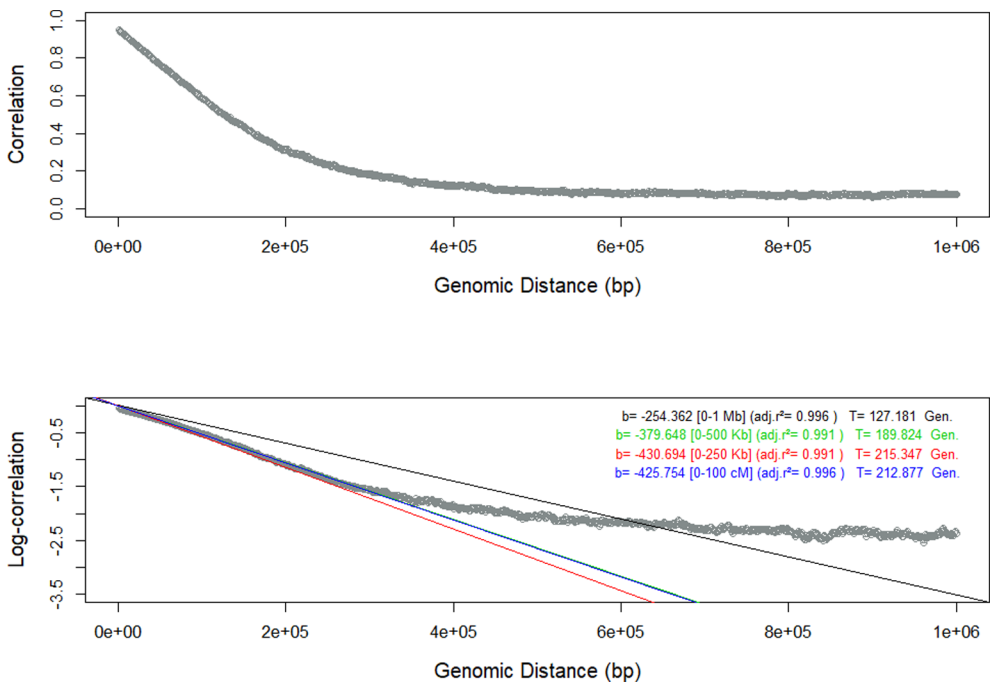


FIGURE 1 | Persistency of Linkage Disequilibrium (LD) measured as the correlation and the log-correlation between the estimates of LD in Retinta (Re) and Rubia Gallega (RG) populations, and estimates of the slope of the regression with respect to the genomic distance for different ranges [0-1 Mb – black, 0-500 Kb-green, 0-250 Kb- red, and 0-100 Kb-blue].

TABLE 2 | Estimated number of generations of divergence between seven beef cattle populations in Spain [Asturiana de los Valles (AV), Avileña - Negra Ibérica (ANI), Bruna dels Pirineus (BP), Morucha (Mo), Pirenaica (P), Retinta (Re), Rubia Gallega (RG)] based on the architecture of Linkage Disequilibrium.

	ANI	BP	Mo	Pi	Re	RG
AV	181.2	132.3	160.7	184.1	185.9	157.2
ANI	–	244.9	133.6	268.6	175.1	225.1
BP	–	–	232.8	168.1	258.1	176.4
Mo	–	–	–	252.8	168.8	205.2
Pi	–	–	–	–	281.9	215.3
Re	–	–	–	–	–	229.1

persistency of LD phase on the genomic distance to -2*c*T, but the analysis was restricted to the SNP markers located within a sliding window of 1 Mb (500 kb downstream and 500 kb upstream) in steps of 0.1 Mb. We obtained as many as 25,098 estimates of *c* for each pair of populations (**Figures S21** to **S41**). The rationale to this approach was that genetic drift and variations on the effective size of the populations should have affected the entire autosomal genome with the same intensity and, therefore, regional variation in the persistency of LD phase should reflect variations in the recombination rate. However, as in other LD-based procedures for estimating the recombination rate (Li and Stephens, 2003), it could also reflect differences in the intensity of the mutation rate or the occurrence of selection events.

Once the local estimates of *c* for the 21 pair of populations were available, we calculated the correlations between them (**Figure 2**). Values ranged from moderate (0.37) to strong (0.76), with an average of 0.55 and a standard deviation of 0.08. The results were consistent with the output of **Table 2**. The minimum correlation (0.37) was obtained between ANI-Mo and BP-Pi, which is consistent with the large number of generations of divergence between ANI and BP (244.9), ANI and Pi (268.6), Mo and BP (232.8), and Mo and Pi (252.8). In contrast, the

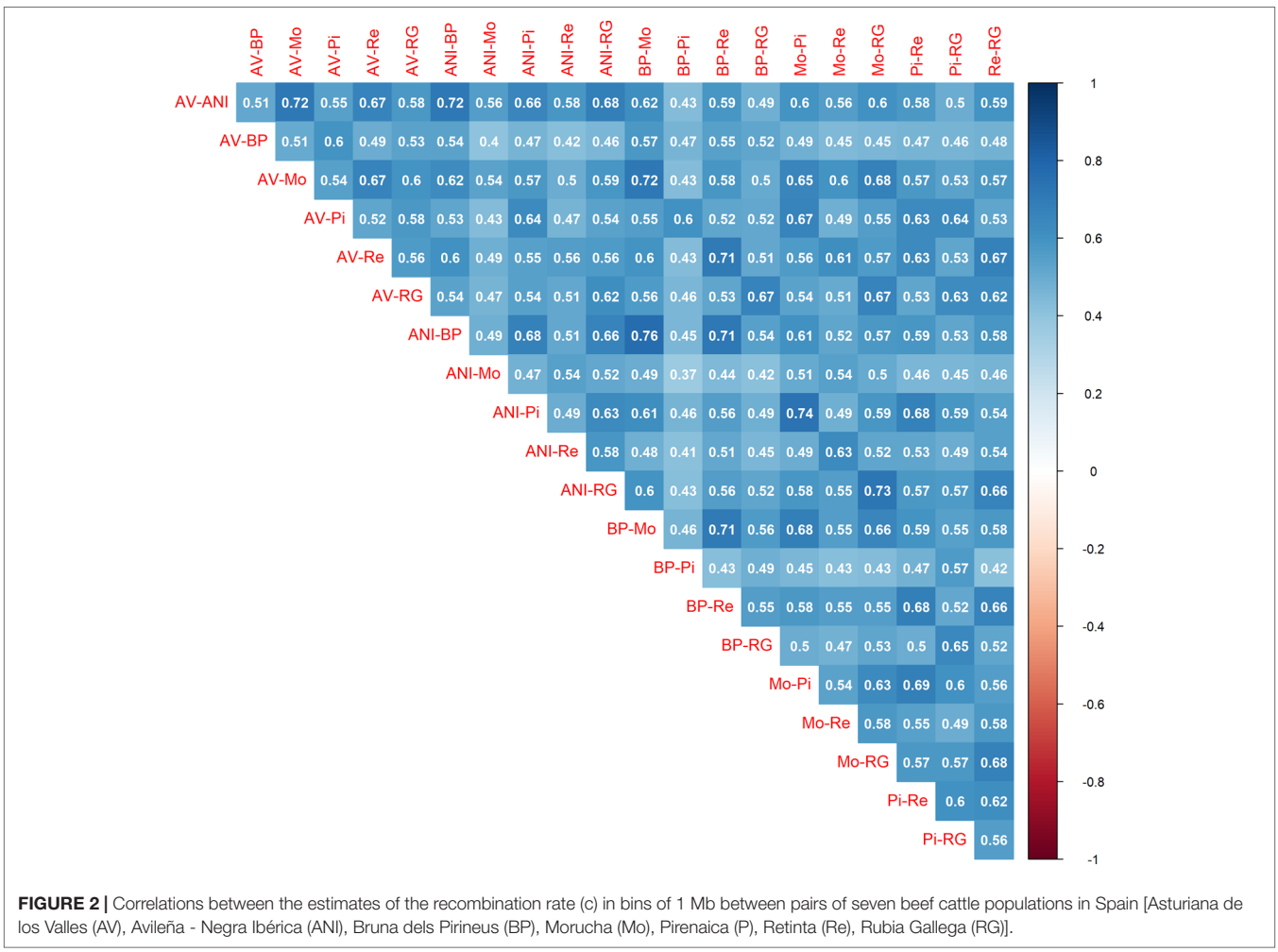


FIGURE 2 | Correlations between the estimates of the recombination rate (c) in bins of 1 Mb between pairs of seven beef cattle populations in Spain [Asturiana de los Valles (AV), Avileña - Negra Ibérica (ANI), Bruna dels Pirineus (BP), Morucha (Mo), Pirenaica (P), Retinta (Re), Rubia Gallega (RG)].

greatest correlation was between ANI-BP and Mo-BP, given that the number of generations of divergence between ANI and Mo is only 133.6 generations. Therefore, given the noise in the LD estimates with small sample sizes, this average correlation and the consistency of estimates between pairs of populations should be considered very relevant. The similarity between estimates was within the same range than the reported between estimates of the recombination rate in human (Graffelman et al., 2007; Laayouni et al., 2011; Manu et al., 2018) and livestock (Petit et al., 2017; Shen et al., 2018) populations. Thus, somehow it confirms that similarities in the distribution of the recombination rate are achieved between closely related populations (Smukowski and Noor, 2011), such as the analyzed in this study (Cañas-Álvarez et al., 2015; González-Rodríguez et al., 2017). This similarity was even observed between pairs of populations that do not share any population (i. e. ANI-AV and BP-Pi), whose average correlation (0.52 ± 0.06) was only slightly lower than the average of the correlations between estimates from pairs that share (i.e. ANI-AV vs ANI-Mo) (0.58 ± 0.08). It reinforces the hypothesis that the similarity between persistence of LD phase at different locations of the genome and between pairs of populations is related with variations in the recombination rate throughout the genome.

Despite divergences associated with a specific population pair and possible selection events, the distribution of the persistency of LD phase appeared to follow a global pattern. Therefore, we used the estimates of *c* to describe the distribution of the recombination rate along the autosomal genome. Initially, we calculated an average of all of the estimates of *c* for each chromosome (**Figure 3**), which ranged from 1.12 cM per Mb in BTA9 to 1.50 cM per Mb in BTA25. In general, the largest chromosomes tended to have the lowest recombination rates. The relationship between recombination rate and chromosome length (Kaback et al., 1992; Jensen-Seaman et al., 2004; Li and Freudenberg, 2009) or genome length (Lynch, 2006) has been reported in several species and may be associated to the difficulties of small chromosomes to find their homologues during meiosis (Tiemann-Boege et al., 2017). In fact, Fledel-Alon et al. (2009) suggested that, in meiosis, each chromosome usually undergoes at least one crossing-over, which produces a very strong correlation between the average number of crossovers and chromosome length.

Next, we evaluated the relationship between the recombination rate and the relative physical position within the chromosome by averaging the *c* estimates for each percentile along the length of each chromosome (**Figure 4**). The results were similar to those of Sandor et al. (2012); Ma et al. (2015), and Shen et al. (2018) in the bovine specie and to the results of haplotype diversity measured in the same populations (Mouresan et al., 2017). In cattle, all autosomal chromosomes are acrocentric (Popescu, 1990) and, in our study, the recombination rate was lowest at the beginning of the chromosome, near the centromere. Furthermore, a low recombination rate was evident at the middle of the chromosome, although the centromere of chromosomes in cattle is not located there. Ma et al. (2015) argued that the bimodal distribution of recombination rates might be caused by positive crossover interference. The highest recombination rate was at the distal portion of the chromosome (over the 95 % percentile of the relative position within chromosomes), in agreement with studies that have shown that recombination rate is highest at the telomeres (Nachman, 2002; Coop and Przeworski, 2007).

Additionally, a map of the recombination rates throughout the genome was calculated by averaging the estimates from the 21 pairs of populations in 0.1 Mb steps along the autosomal chromosomes (**Figure 5**). The average estimate of the recombination rate was 1.275 cM per Mb with a standard deviation of 0.381. As expected, the recombination rate was very similar to the rate assumed in the initial step of the study (1.25 cM per Mb). The estimated recombination rates within the genome were highly

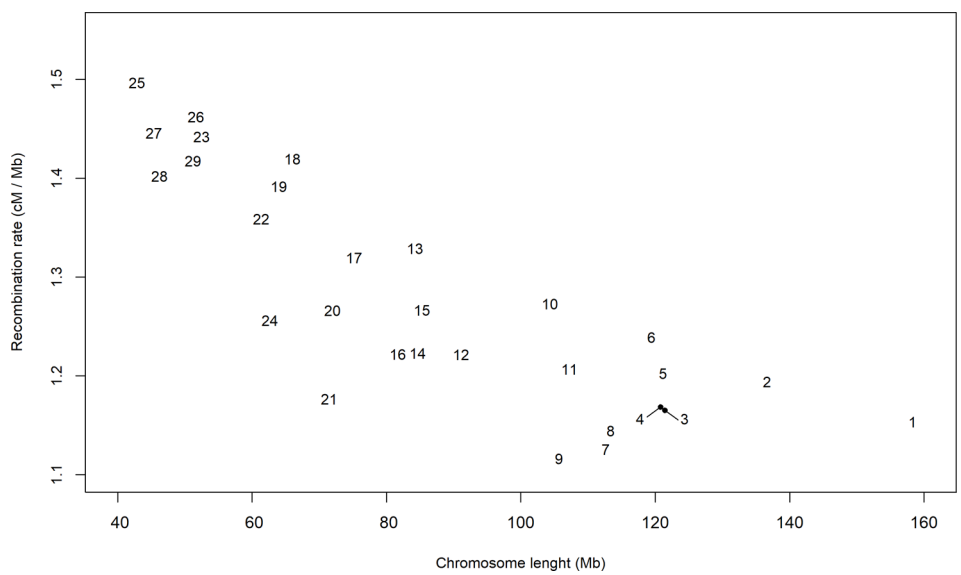


FIGURE 3 | Average estimate per chromosome of the recombination rate (cM/Mb) in bins of 1 Mb and for all pairs of seven breeds of beef cattle populations in Spain.

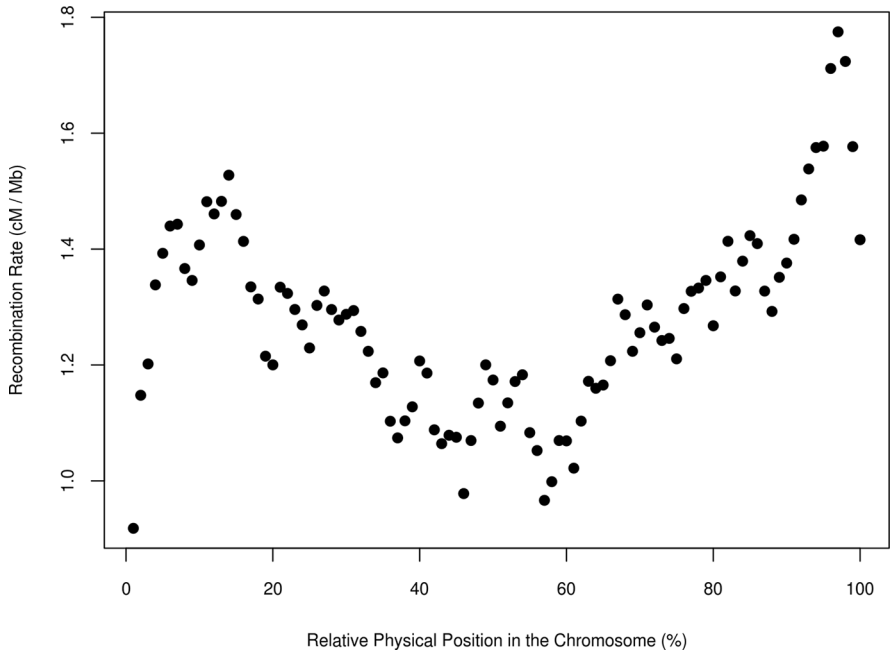


FIGURE 4 | Average estimate of the recombination rate (cM/Mb) in bins of 1 Mb and for all pairs of populations and the relative physical position in the chromosome in seven breeds of beef cattle in Spain.

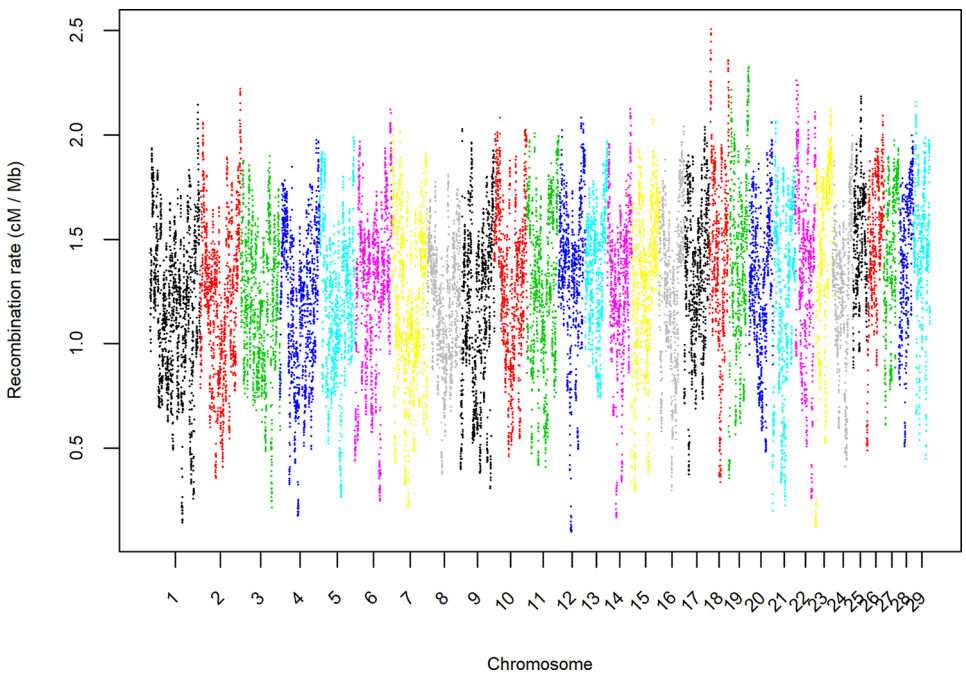


FIGURE 5 | Average estimate of the recombination rate (cM/Mb) in bins of 1 Mb and throughout the autosomal genome for all pairs of seven beef cattle populations in Spain.

heterogeneous, and the ratio between the genomic regions that had the highest (BTA18-7.1 Mb) and the lowest (BTA12-42.4 Mb) estimated recombination rates was 25.37. To illustrate the differences between those two extreme regions of the genome, **Figure 6** displays the average recombination rate and all specific persistencies of LD phase for each population pair.

The heterogeneity of the recombination rate reflected the presence of highly recombining genomic regions. Most of

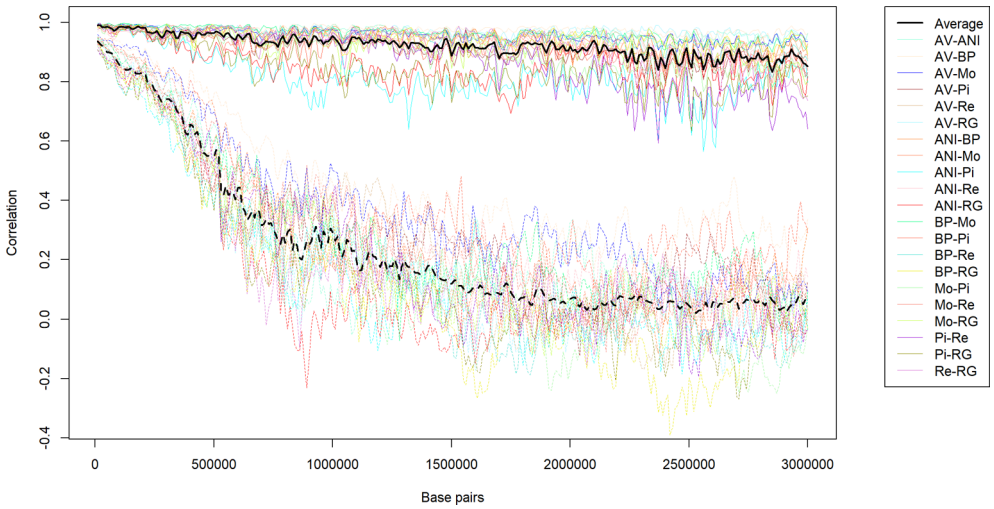


FIGURE 6 | Average (bold lines) and specific population pair (transparent lines) persistency of LD (correlation between LD estimates in pairs of populations) in the genomic regions that had the highest (BTA12-42.4 Mb, continuous lines) or the lowest persistency (BTA18-7,1 Mb, dotted lines) among seven beef breeds in Spain.

the recombination events may occur in a small portion of the genome, as observed in other species. However, the Gini index between the cumulative distributions of the recombination rate and the genetic distance was 0.1803, which is lower than others reported in human (Kong et al., 2002), apes (Stevison et al., 2016), and livestock (Petit et al., 2017) populations. It is likely that the Gini index was low because the method used in our study was only able to distinguish among the rates of recombination (as a measure of the persistency in LD phase) within relatively large genomic regions (1 Mb), and recombination hotspots often are restricted to 1–2 kb (Myers et al., 2005; Mancera et al., 2008).

The 5th and 95th percentiles of the recombination rate estimates were 0.593 and 1.856 cM per Mb, respectively, and we identified the genes within the genomic regions that had values that were either above or below those percentiles. The number of genes within the regions that were above the 95th (high recombination rate – HRR-) or below the 5th (low recombination rate –LRR-) percentiles were 665 and 669, respectively. Some studies have suggested that there is a negative correlation between gene density and the frequency of recombination hotspots (Myers et al., 2005; Freudenberg et al., 2009; Stapley et al., 2017), which was not detectable with the methods used in our study.

Furthermore, some authors (The International HapMap Consortium, 2007) have suggested that genes with highly conserved function are located surrounding regions with low recombination rate; on the other hand, HRR regions contains genes that are exposed to recurrent adaptive process to allow plasticity of organism to coming circumstances. In our study, we have tried to corroborate these statements using ORA with the GO terms for biological processes, cellular components, and molecular functions. The results are presented in **Tables 3** to **5** (*Homo sapiens* annotation database) and in **Supplementary Tables 1** to **3** (*Bos taurus* annotation database). In general, the results with the *Homo sapiens* database yielded results with lower FDR than with the *Bos taurus* database, probably because the human genome in notably more annotated than the bovine one.

TABLE 3 | False discovery rate (FDR) for the top 10 enriched Gene Ontology (GO) terms for Biological processes with the Homo sapiens database for genes within the genomic regions located over the 95th (high recombination rate) and below the 5th (low recombination rate) percentiles of the average recombination rates.

High Recombination Rate		Low Recombination Rate	
GO TERM	**FDR**	**GO TERM**	**FDR**
Protein citrullination	2.00e-04	Homophilic cell adhesion via plasma membrane adhesion molecules	2.45e-01
Histone citrullination	2.00e-04	DNA replication initiation	2.45e-01
Extracellular vesicle biogenesis	3.09e-01	Vitamin A metabolic process	2.45e-01
Regulation of action potential	3.27-01	Cell-cell adhesion via plasma-membrane adhesion molecules	5.84e-01
Regulation of substrate adhesion-dependent cell spreading	3.27-01	Regulation of chemokine biosynthetic process	5.84e-01
Sodium ion transmembrane transport	3.27-01	Celullar response to electrical stimulus	5.84e-01
Peptidyl-arginine modification	3.27-01	Chemokine biosynthetic process	5.84e-01
phagosome acidification	3.27-01	Chemokine metabolic process	5.84e-01
Cardiac muscle cell action potential	3.27-01	Neutrophil mediated killing of bacterium	1.00e-00
Regulation of fibroblast growth factor receptor signaling pathway	3.27-01	Sulfur amino acid catabolic process	1.00e-00

TABLE 4 | False discovery rate (FDR) for the top 10 enriched Gene Ontology (GO) terms for Cellular Components with the Homo sapiens database for genes within the genomic regions located over the 95th (high recombination rate) and below the 5th (low recombination rate) percentiles of average recombination rates.

High Recombination rate		Low Recombination Rate	
GO TERM	**FDR**	**GO TERM**	**FDR**
Neuron part	0.02e-04	Plasma membrane region	2.00e-04
Synapse	0.02e-04	Golgi apparatus	2.70e-03
Synapse part	1.5e-03	Intrinsic component of the plasma membrane	1.25-02
Golgi apparatus	1.5e-03	Cytoplasmic vesicle part	2.83e-02
Cell projection part	1.5e-03	Integral component of the plasma membrane	2.83e-02
Plasma membrane bounded cell projection part	1.5e-03	Perinuclear region of cytoplasm	2.83e-02
Postsynapse	3.0e-03	Cell projection part	3.05e-02
Neuron projection	3.2e-03	Plasma membrane bounded cell projection part	3.05e-02
Vacuole	3.7e-03	Golgi subcompartment	3.05e-02
Phagocytic vesicle	4.8e-03	Plasma membrane protein complex	4.6e-02

TABLE 5 | False discovery rate (FDR) for the top 10 enriched Gene Ontology (GO) terms for Molecular Functions for genes with the Homo sapiens database within the genomic regions located over the 95th (high recombination rate) and below the 5th (low recombination rate) percentiles of the average recombination rates.

High Recombination Rate		High Recombination Rate	
GO TERM	**FDR**	**GO TERM**	**FDR**
Protein-arginine deaminase activity	1.00e-04	GTP-dependent protein binding	8.01-0e2
Hydrolase activity, acting on carbon nitrogen (but not peptide) bonds, in linear amidines	1.13e-02	Aminoacyl-tRNA ligase activity	4.76e-01
Solute: cation antiporter activity	4.74e-02	Ligase activity, forming carbon-oxigen bonds	4.76e-01
Cation: cation antiporter activity	7.49e-02	Molecular carrier activity	4.76e-01
Potassium: proton antiporter activity	1.01e-01	Drug binding	4.76e-01
Metal ion transmembrane transported activity	1.11e-01	Phosphatidylinositol bisphosphate kinase activity	4.76e-01
Solute: Proton antiporter activity	1.16e-01	ARF guanyl-nucleotide exchange factor activity	4.76e-01
Monovalent cation: proton antiporter activity	1.16e-01	Nucleocytoplasmic carrier activity	4.76e-01
Sodium: proton antiporter activity	1.16e-01	Identical protein binding	4.76e-01
Monovalent inorganic cation transmembrane transported activity	1.19e-01	Metalloexopeptidase activity	4.76e-01

The results of the enrichment analysis for biological processes with the human database (**Table 3**) only provided enriched GO terms with a False Discovery Rate (FDR) lower than 0.05 with the genes present in the HRR genomic regions. The significant GO terms correspond to *Protein citrullination* and *Histone citrullination*. Citrullination, the conversion of the amino acid arginine in a protein into the amino acid citrulline, has been related to an increase in antigenic diversity (Nguyen and James, 2016). The higher recombination rate of those genomic regions might indicate that they have evolved to have high plasticity to adapt to changing environments (Charlesworth et al., 2009; Campos et al., 2014). Therefore, the generation of new genetic variants by recombination may help the antigen diversity from the perspective of the host. Thus, it works as a mechanism to adapt its immune response to fight against the ability of the pathogen to modify its antigenic targets. The results obtained from the *Bos taurus* database (**Supplementary Table 1**) were not significant (FDR < 0.05).

The top 10 enriched GO terms for cellular components for the genomic regions that had either high or low recombination rate are presented in **Table 4** (*Homo sapiens*) and **Supplementary Table 2** (*Bos taurus*). For *Homo sapiens*, the FDR was generally lower than it was for biological processes (FDR < 4.6e-02). Some cellular components (*Golgi apparatus*, *Cell projection part, Plasma membrane bounded bell projection part*) occurred in both types of genomic regions, but there were some important differences between them. The HRR genomic regions were enriched with genes whose expression is located at the extracellular space and related with neuronal interactions (*neuron part, synapse, synapse part, postsynapse, or neuron projection*). In contrast, the genes located at LRR regions were associated with very basic intracellular (*Cytoplasmic vesicle part*, *Perinuclear region of the cytoplasm*) or membrane components (*Plasma membrane region*, *Intrinsic component of the plasma membrane Integral component of the plasma membrane*, *Plasma membrane protein complex*). The results obtained from the *Bos taurus* database were significant (FDR < 0.05) only for *cytosol* and *nuclear lumen* in the LRR regions, confirming the results provided by the *Homo sapiens* database.

The enriched GO terms for molecular functions are presented in **Table 5** (*Homo sapiens*) and **Supplementary Table 3** *(Bos Taurus*). The three significantly (FDR < 0.05) enriched GO terms for the HRR genomic regions were coherent with the enriched GO terms for biological processes. In fact, two of those were clearly associated with citrullination [*Protein-arginine deiminase activity* and *Hydrolase activity, acting on carbon nitrogen (but not peptide) bonds, in linear amidines*] and the other (*Solute: cation antiporter activity)* was linked to transmembrane transportation of solutes. In contrast, the only significantly enriched GO terms for LRR genomic regions was *GTP-dependent protein binding* (FDR = 8.01e-02), which was confirmed with the results from the *Bos taurus* database (FDR = 2.7e-02). The genes located in the LRR genomic regions should be more conserved, since they may be

necessary for basic functions of the organism. In this sense, the genes belonging to the *GTP-dependent protein GO term* regulate guanine nucleotide-binding proteins that play a crucial role in signal transduction and in a large number of cellular processes (Zachariou et al., 2012).

The results of this study confirm that the genomic architecture of persistency of LD phase is well conserved among closely related populations, such as the Spanish autochthonous beef cattle breeds, and is heterogeneous within the autosomal genome and that this heterogeneity can be used to estimate the recombination rate. Several studies have estimated the persistency of LD phase between populations as a measure of genetic diversity (De Roos et al., 2008; Villa-Angulo et al., 2009; Cañas-Álvarez et al., 2016) and as a mean of predicting the marker density required for multi-breed genomic evaluation (De Roos et al., 2008; Cañas-Álvarez et al., 2016). Nevertheless, the genetic architecture of persistency of phase within the genome has received limited attention. In this study, we estimated the persistency of LD phase among seven beef cattle populations in Spain and its distribution within the genome, which has been related with the genetic architecture of the recombination rates. Even though the recombination rate varies among species, sex, and populations (Dapper and Payseur, 2017), some general patterns were described (Tiemann-Boege et al., 2017). The patterns were confirmed in our analysis by the similitude of our results with some studies in other species or populations (Ma et al., 2015; Shen et al., 2018).

Therefore, the main conclusion of this study is that the heterogeneity of persistency of LD phase between closely related populations can be used to estimate the recombination rate with the procedure developed here, which is simpler and brings similar results as more complex and coalescent dependent methods. This implies that it may help to identify regions related to hot and cold spots when data from several populations of the same ancestral origin are available. Nevertheless, our procedure has several limitations since differences in the mutation rate or selection events can locally affect the persistency of LD phase and it requires populations to be close enough that the recombination rate is well conserved.

AUTHOR CONTRIBUTIONS

The study was conceived by LV. Preliminary data preparation was done by EM, AG-R, JC-Á, and SL. Data analysis was done by EM and LV. EM and LV produced the draft manuscript. JA, CD, AM, and JP collaborated in generating the data. SL, JA, CD, JB, AM, PL-B, and JP discussed the results. All authors reviewed and revised the manuscript.

FUNDING

We thank the AGL 2010-15903 grant from the Spanish government and the KBBE.2011.1.3-06 project from the European Union's Seventh Framework. We also thank the Breed Societies for their collaboration in collecting samples. The support of FEAGAS is acknowledged. JC-Á acknowledges the COLCIENCIAS support by the Francisco José De Caldas Fellowship 497/2009, and AG-R acknowledges the financial support provided by the BES-2011-045434 fellowship.

ACKNOWLEDGMENTS

We thank Bruce Macwirther and María Martínez-Castillero for their help in the language editing. Preliminary results of this study were presented at the World Congress of Genetics Applied to Livestock Production, 11.285.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01170/full#supplementary-material

REFERENCES

Arias, J. A., Keehan, M., Fisher, P., Coppieters, W., and Spelman, R. (2009). A high density linkage map of the bovine genome. *BMC Genet.* 10, 18. doi: 10.1186/1471-2156-10-18

Arnheim, N., Calabrese, P., and Tiemann-Boege, I. (2007). Mammalian meiotic recombination hot spots. *Annu. Rev. Genet.* 41, 369–399. doi: 10.1146/annurev.genet.41.110306.130301

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., et al. (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science* 13, 193–198. doi: 10.1126/science.1216872

Biegelmeyer, P., Gulias-Gomes, C. C., Caetano, A. R., Steibel, J. P., and Cardoso, F. F. (2016). Linkage disequilibrium, persistence of phase and effective population size estimates in hereford and braford cattle. *BMC Genet.* 17 (1), 32. doi: 10.1186/s12863-016-0339-8

Brito, L. F., Jafarikia, M., . Grossi, D. A., Kijas, J. W., . Porto-Neto, L. R., Ventura, R. V., et al. (2015). Characterization of linkage disequilibrium, consistency of gametic phase and admixture in australian and canadian goats. *BMC Genet.* 16 (1), 67. doi: 10.1186/s12863-015-0220-1

Brito, L. F., McEwan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., Dodds, K. G., et al. (2017). Genetic diversity of a new zealand multi-breed sheep population and composite breeds' history revealed by a high-density SNP chip. *BMC Genet.* 18 (1), 25. doi: 10.1186/s12863-017-0492-8

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

Cañas-Álvarez, J. J., González-Rodríguez, A., Munilla, S., Varona, L., Díaz, C., Baro, J. A., et al. (2015). Genetic diversity and divergence among spanish beef cattle breeds assessed by a bovine high-density SNP chip. *J. Anim. Sci.* 93 (11), 5164–5174. doi: 10.2527/jas2015-9271

Cañas-Álvarez, J. J., Mouresan, E. F., Varona, L., Díaz, C., Molina, A., Baro, J. A., et al. (2016). Linkage disequilibrium, persistence of phase, and effective population size in spanish local beef cattle breeds assessed through a high-density single nucleotide polymorphism chip. *J. Anim. Sci.* 94 (7), 2779–2788. doi: 10.2527/jas2016-0425

Campos, J. L., Halligan, D. L., Haddrill, P. R., and Charlesworth, B. (2014). The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. *Mol. Biol. Evol.* 31 (4), 1010–1028. doi: 10.1093/molbev/msu056

Ceriani, L., and Verme, P. (2012). The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *J. Economic Inequality* 10 (3), 421–443. doi: 10.1007/s10888-011-9188-x

Charlesworth, B., Betancourt, A. J., Kaiser, V. B., and Gordo, I. (2009). Genetic recombination and molecular evolution. *Cold Spring Harb. Symp. Quant. Biol.* 74, 177–186. doi: 10.1101/sqb.2009.74.015

Coop, G., and Przeworski, M. (2007). An evolutionary view of human recombination. *Nat. Rev. Genet.* 8 (1), 23–34. doi: 10.1038/nrg1947

Dapper, A. L., and Payseur, B. A. (2017). Connecting theory and data to understand recombination rate evolution. *Philos. Trans. R. Soc. B: Biol. Sci.* 372 (1736), 20160469. doi: 10.1098/rstb.2016.0469

De Roos, A. P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in holstein-friesian, jersey and angus cattle. *Genetics* 179, 1503–1512. doi: 10.1534/genetics.107.084301

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Harlow, UK: Longman Group.

Fledel-Alon, A., Wilson, D. J., Broman, K., Wen, X., and Ober, C. (2009). Broad-scale recombination patterns underlying proper disjunction in humans. *PloS Genet.* 5 (9), 1000658. doi: 10.1371/journal.pgen.1000658

Flicek, P., Ahmed, I., Ridwan Amode, M., Barrell, D., Beal, K., Brent, S., et al. (2013). Ensembl 2013. *Nucleic Acids Res.* 41, D48–D55. doi: 10.1093/nar/gks1236

Freudenberg, J., Wang, M., Yang, Y., and Li, W. (2009). Partial correlation analysis indicates causal relationships between gc-content, exon density and recombination rate in the human genome. *BMC Bioinf.* 10, S66. doi: 10.1186/1471-2105-10-S1-S66

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science* 296 (5576), 2225–2229. doi: 10.1126/science.1069424

Gil, M., Serra, X., Gispert, M., Oliver, M.A.Á., SanÄ udo, C., Panea, B., et al. (2001). The effect of breed-production systems on the myosin heavy chain 1, the biochemical characteristics and the colour variables of longissimus thoracis from seven spanish beef cattle breeds. *Meat Sci.* 58, 181–188. doi: 10.1016/s0309-1740(00)00150-9

González-Rodríguez, A., Munilla, S., Mouresan, E. F., Cañas-Álvarez, J. J., Díaz, C., Piedrafita, J., et al. (2016). On the performance of tests for the detection of signatures of selection: a case study with the Spanish autochthonous beef cattle populations. *Genet. Selection Evol.* 48 (1). doi: 10.1186/s12711-016-0258-1

González-Rodríguez, A., Munilla, S., Mouresan, E. F., Cañas-Álvarez, J. J., Baro, J. A., Molina, A., et al. (2017). Genomic differentiation between asturiana de los valles, avileña-negra ibérica, bruna dels pirineus, morucha, pirenaica, retinta and rubia gallega cattle breeds. *Animal* 11 (10), 1667–1679. doi: 10.1017/S1751731117000398

Graffelman, J., Balding, D. J., Gonzalez-Neira, A., and Bertranpetit, J. (2007). Variation in estimated recombination rates across human populations. *Hum. Genet.* 122 (3–4), 301–310. doi: 10.1007/s00439-007-0391-6

Grossi, D. A., Jafarikia, M., Brito, L. F., Buzanskas, M. E., Sargolzaei, M., and Schenkel, F. S. (2017). Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in canadian pigs. *BMC Genet.* 18 (1), 6. doi: 10.1186/s12863-017-0473-y

Guryev, V., Smits, B. M. G., Belt, J. V. D., Verheul, M., Hubner, N., and Cuppen, E. (2006). Haplotype block structure is conserved across mammals. *PloS Genet.* 7, 1111–1118. doi: 10.1371/journal.pgen.0020121

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* 13 (4), 635–643. doi: 10.1101/gr.387103

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38 (6), 226–231. doi: 10.1007/BF01245622

Hinrichs, A. L., Larkin, E. K., and Suarez, B. K. (2009). Population stratification and patterns of linkage disequilibrium. *Genet. Epidemiol.* 33, 88–92. doi: 10.1002/gepi.20478

Hulten, M. A., Palmer, R. W., and Laurie, D. A. (1982). Chiasma derived genetic maps and recombination fractions: chromosome 1. *Ann. Hum. Genet.* 46 (2), 167–175. doi: 10.1111/j.1469-1809.1982.tb00707.x

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., et al. (2004). Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14, 528–538. doi: 10.1101/gr.1970304

Kaback, D., Guacci, V., Barber, D., and Mahon, J. (1992). Chromosome size-dependent control of meiotic recombination. *Science* 256 (5054), 228–232. doi: 10.1126/science.1566070

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., et al. (2002). A high-resolution recombination map of the human genome. *Nat. Genet.* 31 (3), 241–247. doi: 10.1038/ng917

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467 (7319), 1099–1103. doi: 10.1038/nature09525

Laayouni, H., Montanucci, L., Sikora, M., Melé, M., Dall'Olio, G. M., Lorente-Galdos, B., et al. (2011). Similarity in recombination rate estimates highly correlates with genetic differentiation in humans. Edited by Carles Lalueza-Fox. *PloS One* 6 (3), e17913. doi: 10.1371/journal.pone.0017913.

Li, W., and Freudenberg, J. (2009). Two-parameter characterization of chromosome-scale recombination rate. *Genome Res.* 19 (12), 2300–2307. doi: 10.1101/gr.092676.109

Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165 (4), 2213–2233.

Lynch, M. (2006). The origins of eukaryotic gene structure. *Mol. Biol. Evol.* 23 (2), 450–468. doi: 10.1093/molbev/msj050

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., et al. (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PloS Genet.* 11 (11), e1005387. doi: 10.1371/journal.pgen.1005387. Edited by Adam J. Auton.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454 (7203), 479–485. doi: 10.1038/nature07135

Manu, S., Acharya, K. K., and Thiyagarajan, S. (2018). Systematic analyses of autosomal recombination rates from the 1000 genomes project uncovers the global recombination landscape in humans. *BioRxiv*, 246702. doi: 10.1101/246702

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160 (3), 1231–1241.

Ministerio de Medio Ambiente y Medio Rural y Marino (2010). *Razas de Ganado Del Catálogo Oficial En España*. Madrid, Spain: Ministerio de Medio Ambiente y Medio Rural y Marino.

Mokry, F., Buzanskas, M., Mudadu, M. A., Grossi, D. A., Higa, R., Ventura, R., et al. (2014). Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. *BMC Genomics* 15 (Suppl 7), S6. doi: 10.1186/1471-2164-15-S7-S6

Mouresan, E. F., González-Rodríguez, A., Cañas-Álvarez, J. J., Díaz, C., Altarriba, J., Baro, J. A., et al. (2017). On the haplotype diversity along the genome in spanish beef cattle populations. *Livestock Sci.* 201, 33–33. doi: 10.1016/j.livsci.2017.04.015

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310 (5746), 321–324. doi: 10.1126/science.1117196

Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Curr. Opin. Genet. Dev.* 12 (6), 657–663. doi: 10.1016/s0959-437x(02)00358-1

Nguyen, H., and James, E. A. (2016). Immune recognition of citrullinated epitopes. *Immunology* 149 (2), 131–138. doi: 10.1111/imm.12640

Paigen, K., and Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nat. Rev. Genet.* 11 (3), 221–233. doi: 10.1038/nrg2712

Park, L. (2012). Linkage disequilibrium decay and past population history in the human genome. *PloS One* 7 (10), e46603. doi: 10.1371/journal.pone.0046603. Edited by Thomas Mailund.

Petit, M., Astruc, J.-M., Sarry, J., Drouilhet, L., Fabre, S., Moreno, C., et al. (2017). Variation in recombination rate and its genetic determinism in sheep populations. *Genetics* 207 (2). doi: 10.1534/genetics.117.300123

Piedrafita, J., Quintanilla, R., Sañudo, C., Olleta, J.-L., Campo, M.-M., Panea, B., et al. (2003). Carcass quality of 10 beef cattle breeds of the southwest of europe in their typical production systems. *Livestock Prod. Sci.* 82 (1), 1–13. doi: 10.1016/S0301-6226(03)00006-X

Popescu, P. C. (1990). Chromosomes of the cow and the bull. *Adv. Vet. Sci. Comp. Med.* 34, 41–71. doi: 10.1016/B978-0-12-039234-6.50007-0

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., and Georges, M. (2012). Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. Edited by Kenneth Paigen. *PloS Genet.* 8 (7), e1002854. doi: 10.1371/journal.pgen.1002854

Sarbajna, S., Denniff, M., Jeffreys, A. J., Neumann, R., Artigas, M. S., Veselis, A., et al. (2012). A major recombination hotspot in the xqyq pseudoautosomal region gives new insight into processing of human gene conversion events. *Hum. Mol. Genet.* 21 (9), 2029–2038. doi: 10.1093/hmg/dds019

Shen, B., Jiang, J., Seroussi, E., Liu, G. E., and Ma, L. (2018). Characterization of recombination features and the genetic basis in multiple cattle breeds. *BMC Genomics* 19 (1), 1–10. doi: 10.1186/s12864-018-4705-y

Smukowski, C. S., and Noor, M. A. F. (2011). Recombination rate variation in closely related species. *Heredity* 107 (6), 496–508. doi: 10.1038/hdy.2011.44

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., and Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philos. Trans. R. Soc. B: Biol. Sci.* 372 (1736), 20160455. doi: 10.1098/rstb.2016.0455

Stevison, L. S., Woerner, A. E., Kidd, J. M., Kelley, J. L., Veeramah, K. R., McManus, K. F., et al. (2016). The time scale of recombination rate evolution in great apes. *Mol. Biol. Evol.* 33 (4), 928–945. doi: 10.1093/molbev/msv331

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *J. Exp. Zool.* 14 (1), 43–59. doi: 10.1002/jez.1400140104

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., et al. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17 (4), 520–526. doi: 10.1101/gr.6023607

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449 (7164), 851–861. doi: 10.1038/nature06258

Tiemann-Boege, I., Schwarz, T., Striedner, Y., and Heissl, A. (2017). The consequences of sequence erosion in the evolution of recombination hotspots. *Philos. Trans. R. Soc. B: Biol. Sci.* 372 (1736), 20160462. doi: 10.1098/rstb.2016.0462

Villa-Angulo, R., Matukumalli, L. K., Gill, C. A., Choi, J., Tassell, C. P. V., and Grefenstette, J. J. (2009). High-resolution haplotype block structure in the cattle genome. *BMC Genet.* 10 (1), 19. doi: doi.org/10.1186/1471-2156-10-19

Wall, J. D., and Stevison, L. S. (2016). Detecting recombination hotspots from patterns of linkage disequilibrium. *G3 (Bethesda)* 6 (8), 2265–2271. doi: 10.1534/g3.116.029587

Xu, L., Zhu, B., Wang, Z., Xu, L., Liu, Y., Chen, Y., et al. (2019). Evaluation of linkage disequilibrium, effective population size and haplotype block structure in chinese cattle. *Animals* 9 (3), 83. doi: 10.3390/ani9030083

Zachariou, V., Duman, R. S., and Nestler, E. (2012). "G proteins," in *Basic neurochemistry.* Eds. Brady, S. T., Siegel, G. J., Albers, R. W., and Price, D. L. (Waltham, MA. USA: Academic Press), 411–422. doi: 10.1016/B978-0-12-374947-5.00021-3

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 December 2019
doi: 10.3389/fgene.2019.01269

Population Structure and Genetic Diversity of Nile Tilapia (*Oreochromis niloticus*) Strains Cultured in Tanzania

*Redempta A. Kajungiro[1,2], Christos Palaiokostas[1], Fernando A. Lopes Pinto[1], Aviti J. Mmochi[3], Marten Mtolera[3], Ross D. Houston[4] and Dirk Jan de Koning[1]**

[1] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden, [2] Department of Aquatic Science and Fisheries, College of Agricultural Sciences and Fisheries Technology, University of Dar es Salaam, Dar es Salaam, Tanzania, [3] Institute of Marine Sciences, University of Dar es Salaam, Dar es Salaam, Tanzania, [4] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom

OPEN ACCESS

Edited by:
Farai Catherine Muchadeyi, Agricultural Research Council of South Africa (ARC-SA), South Africa

Reviewed by:
Jesús Fernández, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria INIA, Spain
Charles Hefer, AgResearch, New Zealand

***Correspondence:**
Dirk Jan de Koning
dj.de-koning@slu.se

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 08 February 2019
Accepted: 18 November 2019
Published: 20 December 2019

Citation:
Kajungiro RA, Palaiokostas C, Pinto FAL, Mmochi AJ, Mtolera M, Houston RD and de Koning DJ (2019) Population Structure and Genetic Diversity of Nile Tilapia (Oreochromis niloticus) Strains Cultured in Tanzania. Front. Genet. 10:1269. doi: 10.3389/fgene.2019.01269

Understanding population structure and genetic diversity within and between local Nile tilapia lines cultured in Tanzania is important for sustainable aquaculture production. This study investigated the genetic structure and diversity among seven Nile tilapia populations in Tanzania (Karanga, Igunga, Ruhila, Fisheries Education and Training Agency, Tanzania Fisheries Research Institute, Kunduchi, and Lake Victoria). Double-digest restriction site-associated DNA (ddRAD) libraries were prepared from 140 individual fish (20 per population) and sequenced using an Illumina HiSeq 4000 resulting in the identification of 2,180 informative single nucleotide polymorphisms (SNPs). Pairwise F_{st} values revealed strong genetic differentiation between the closely related populations; FETA, Lake Victoria, and Igunga and those from TAFIRI and Karanga with values ranging between 0.45 and 0.55. Population structure was further evaluated using Bayesian model-based clustering (STRUCTURE) and discriminant analysis of principal components (DAPC). Admixture was detected among Karanga, Kunduchi, and Ruhila populations. A cross-validation approach (25% of individual fish from each population was considered of unknown origin) was conducted in order to test the efficiency of the SNP markers to correctly assign individual fish to the population of origin. The cross-validation procedure was repeated 10 times resulting in 77% of the tested individual fish being allocated to the correct population. Overall our results provide a new database of informative SNP markers for both conservation management and aquaculture activities of Nile tilapia strains in Tanzania.

Keywords: aquaculture, population structure, genetic diversity, Nile tilapia, double-digest restriction site-associated DNA-sequencing

INTRODUCTION

Tanzania is a diversity hotspot of tilapias including more than 30 *Oreochromis* species of which 10 are only found in the country (Genner et al., 2018; Shechonge et al., 2019). *Oreochromis niloticus* is the most widespread tilapiine cichlid both in Tanzania and worldwide. During the last 5 years, Nile tilapia aquaculture in Tanzania has increased from 958 MT in 2011 to 4080 MT in 2017 (Kajungiro et al. unpublished data) with a continuously increasing demand for further expansion. Despite the

interest and potential of tilapia aquaculture to contribute to local food production, currently no selective breeding program exists in Tanzania—a situation typical of many African nations.

Common hatchery aquaculture practices could result in a rapid reduction of the genetic diversity of the farmed animals. A well-managed breeding program on the other hand would enable cumulative genetic improvement of target traits, while simultaneously minimize inbreeding and loss of diversity. Forming a base population containing high genetic diversity will be crucial for the success of any future breeding program in Tanzania (Fernández et al., 2014; García-Ballesteros et al., 2017). Furthermore, introductions of fish from one region to another have affected the genetic diversity and population structure of many teleost fish species (Basiita et al., 2018). Due to mismanagement and uncontrolled movement of fish from different regions there is limited information relating to the genetic structure of Nile tilapia strains and their distribution in Tanzania.

Tilapia species have a very complex genetic structure, in common with many other Cichlid fish species (Bezault et al., 2011). Moreover, hybridization and introgression are fairly common in tilapias constituting the management of both wild and farmed populations particularly challenging (Shirak et al., 2009; Wu and Yang, 2012). The aforementioned issue is further exacerbated by the common situation of reproductive viable hybrids in tilapias (Wohlfarth and Hulata, 1982). In addition, ecological factors such as environmental heterogeneity and geographic connectivity have shaped the current population structure and distribution of Nile tilapia in Africa (Bezault et al., 2011).

Genetic diversity plays a crucial role in the adaptation ability of a population in the face of fluctuating environmental conditions (Markert et al., 2010). Conservation programs aim to minimize the loss of genetic diversity in order to increase the chances of successful population restoration and long-term viability. Translocation of fish to supplement suppressed populations may have in fact harmful effects if the recipient population is genetically different (Allendorf and Luikart, 2007). Available knowledge regarding the genetic diversity of cultured strains can also assist in genetic improvement, rearing management and performance potential in various culture environments (Angienda et al., 2011). Further, in selective breeding programs the genetic diversity between and within breeds and populations can provide valuable information regarding the potential response to selection (Oldenbroek, 2017). Due to a high demand from aquaculture, Nile tilapia strains and other unknown tilapia species have been introduced outside their natural geographical distributions in Tanzania (Philippart and Ruwet, 1982; Shechonge et al., 2019). In addition, hybridization with the local tilapia species has been recently reported (Shechonge et al., 2018).

Genetic markers offer a reliable approach for unveiling the genetic structure both among and within populations. In addition, genetic markers can assist in identifying species, individuals or population of origin of unknown samples allowing the authorities in monitoring protected nature reserved areas. As such, knowledge of population genetic structure and genetic diversity of *O. niloticus* is crucial both for conservation practices and for fish breeders. Previous studies examined the genetic structure and diversity between populations of Nile tilapia (*Oreochromis niloticus*), based either on phenotypic traits (Trewavas, 1983), allozymes (Sodsuk and McAndrew, 1991), mitochondrial DNA (Romana-Eguia et al., 2004), randomly amplified polymorphic DNA (Hassanien et al., 2004) or microsatellites (Bhassu et al., 2004; Hassanien and Gilbey, 2005; Mireku et al., 2017). However, the genetic markers used to date have limitations regarding their maximal resolution in detecting the complex genetic structure typically encountered in Nile tilapia populations. Furthermore, to our knowledge no prior study attempted to test the efficiency of genetic markers for predicting the population of origin in putative unknown tilapia samples.

Next-generation sequencing (NGS) technologies have facilitated the discovery of large numbers of genetic markers for practically any organism at an affordable cost allowing the investigation of genetic diversity within and between populations (Candy et al., 2015). Restriction-site associated DNA (RAD) and double-digest RAD (ddRAD) sequencing are NGS-based techniques providing a reduced representation of the studied genome (Baird et al., 2008; Peterson et al., 2012). ddRAD-seq and similar genotyping by sequencing techniques rely on digestion of the genomic DNA with restriction enzyme(s), and subsequent high-depth sequencing of the flanking regions of the cut site. Such genotyping by sequencing techniques have been widely applied in aquaculture species (Robledo et al., 2018). Several studies have applied ddRAD-seq sequencing to generate high-density linkage maps (Brown et al., 2016; Manousaki et al., 2016) and estimate genetic diversity (Antoniou et al., 2017; Hosoya et al., 2018). Furthermore, ddRAD-seq has been utilized in several tilapia studies for evaluating the suitability of DNA from skin mucus swabs (Taslima et al., 2017), identification of sex determining regions (Wessels et al., 2017), and quantitative trait loci (QTL) analysis (Li et al., 2017).

The current study investigated the population genetic structure of seven Nile tilapia populations from Tanzania using ddRAD-seq derived single nucleotide polymorphisms (SNPs). Genetic diversity parameters and population structure using both multivariate analysis and Bayesian clustering algorithms were evaluated. Admixture levels between the different populations were estimated providing valuable information for future management of Nile tilapia resources in Tanzania. Finally, a cross-validation scheme was applied in order to test the efficiency of the generated SNPs for assignment of individual fish to their population of origin.

MATERIALS AND METHODS

Ethics Statement

This study was carried out in accordance with the law on the protection of animals against cruelty (Act no. 12/1974. of the United Republic of Tanzania) upon its approval by the department of Zoology and Wildlife Conservation, University of Dar es salaam. All the permits required to sample wild animals in Tanzania were adhered; these include Research clearance from Tanzania Commission for Science and Technology (COSTECH) and other relevant authorities.

Fish Sample Collection and Preparation

Farmed stocks of *Oreochromis niloticus* juveniles were collected in 2017 from Government aquaculture centers distributed throughout Tanzania. In particular we collected animals from six farmed populations namely: Tanzania Fisheries Research Institute (TAFIRI; −2.5805° S, 32.8979° E), Fisheries Education and Training Agency (FETA; −2.5851° S, 32.8980° E), Karanga (−3.373680° S, 37.318390° E), Igunga (−4.285810° S, 33.879020° E), Kunduchi (−6.670220° S, 39.214840° E), Ruhila (−10.665510° S, 35.645040° E, and one natural population from Lake Victoria (−2.556348° S, 32.881061° E) (**Figure 1**). FETA and TAFIRI are located along Lake Victoria. The TAFIRI stock originated from Lake Victoria in 2014, while the other populations (FETA and Igunga) were stocked in 2016 (personal communication with fish farmer). Igunga is located in the central part of the country, Karanga in the northern part, Kunduchi along the coast of the Indian Ocean, and Ruhila in the southern part of Tanzania (**Figure 1**). Fish were kept in separate hapas (2 m × 2 m) within an earthen pond at Kunduchi Campus for 4 months. Species identification was based on both prior available records for each population and on morphology characteristics as explained by Trewavas (1983): In particular *O. niloticus* were distinguished from other species by large deep-bodied size with relatively small heads and the presence of regular vertical stripes throughout the depth of caudal fin. A total of 140 fish weighing from 50 to 150 g were used in the study. The fish were sedated using pure clove oil at the dosage of 2 ml clove oil to 20 L of water (Fernandes et al., 2017). Twenty fish from each population were fin clipped. Fin clips were stored in 95% ethanol at −20°C, until DNA extraction.

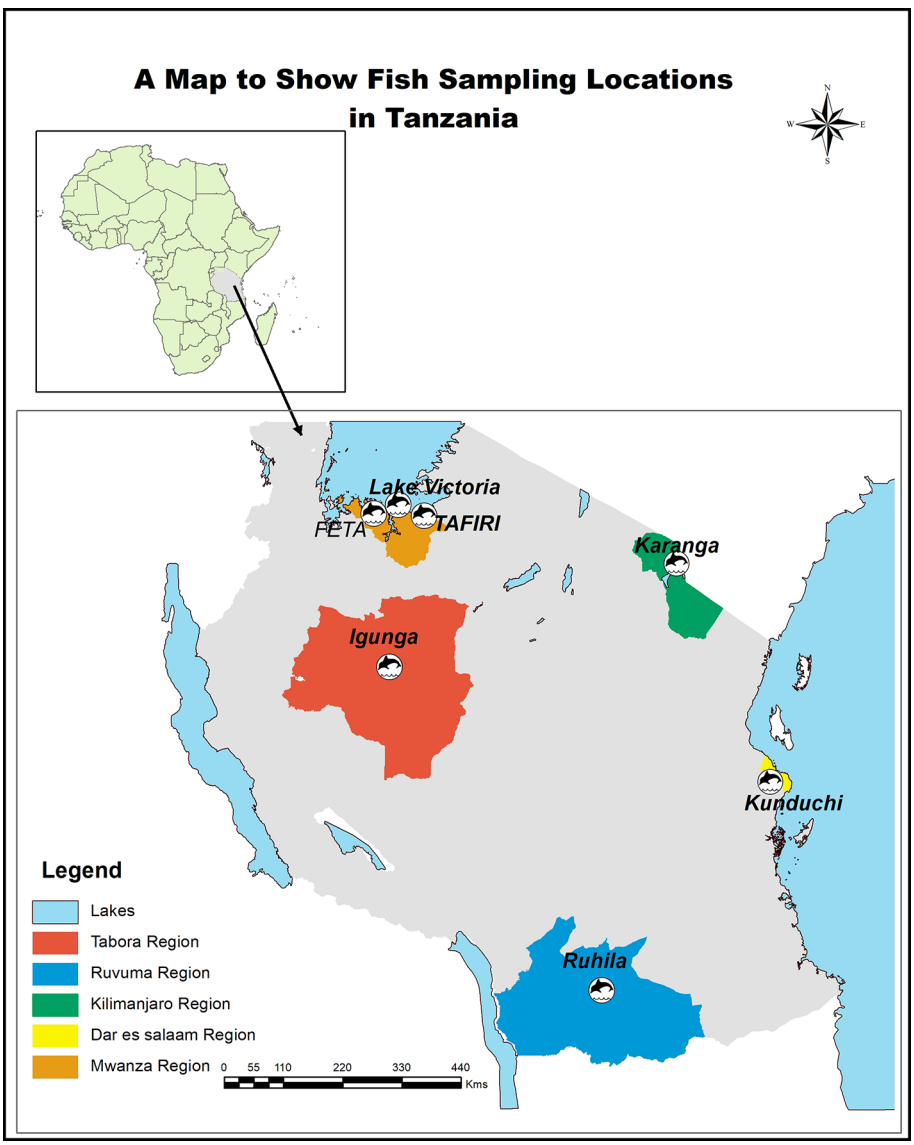


FIGURE 1 | Sampling locations of fish used in the present study.

DNA Extraction

Genomic DNA was extracted from 0.02 g of fish fin using a spin column (QIAsymphony DSP DNA Mini Kit; Qiagen, Hilden, Germany) and eluted into 100 µl of AE (EDTA) buffer (Qiagen) according to the manufacturer's tissue protocol and procedures. The purity and concentration of the extracted DNA were quantified using Qubit 2.0 Fluorometer (Invitrogen). Samples were diluted with Tris EDTA (TE) buffer (Thermo Fisher Scientific) to 25 ng/µl and 2 µl were run on a 1% agarose gel by electrophoresis. Diluted samples were stored at −20°C.

Double-Digest Restriction Site-Associated DNA Library Preparation and Sequencing

ddRAD library preparation was performed according to Peterson et al. (2012), with minor modifications described in Palaiokostas et al. (2015). Briefly, each sample (25 ng DNA) was digested at 37°C for 60 min with *Sbf*I (recognizing the CCTGCA|GG motif) and *Sph*I (recognizing the GCATG|C motif) high fidelity restriction enzymes (New England Biolabs, UK; NEB), using 6 U of each enzyme per microgram of genomic DNA in 1× Reaction Buffer 4 (NEB). The reactions (5 µl final volumes) were then heat inactivated at 65°C for 20 min. Individual-specific combinations of P1 and P2 adapters, each with a unique 5 or 7 bp barcode, were ligated to the digested DNA at 22°C for 60 min by adding 1 µl *Sbf*I compatible P1 adapter (25 nM), 0.7 µl *Sph*I compatible P2 adapter (100 nM), 0.06 µl 100 mmol/L rATP (Promega, UK), 0.95 µl 1× Reaction Buffer 2 (NEB), 0.05 µl T4 ligase (NEB, 2×10^6 U/ml) and reaction volumes made up to 8 µl with nuclease-free water for each sample. Following heat inactivation at 65°C for 20 min, the ligation reactions were slowly cooled to room temperature (over 1 h) then combined in a single pool (for one sequencing lane) and purified. Size selection (300–600 bp) was performed by agarose gel separation and followed by gel purification and PCR amplification. A total of 100 µl of the amplified libraries (13–14 PCR cycles) was purified using an equal volume of AMPure beads. After eluting into 20 µl EB buffer (MinElute Gel Purification Kit, Qiagen, UK), the libraries were ready for sequencing. The libraries were sequenced at Edinburgh Genomics Facility, University of Edinburgh on an Illumina HiSeq 4000 instrument.

Sequence Data Analysis and Single Nucleotide Polymorphism Genotyping

Reads of low quality (Q < 20) and missing the expected restriction sites were discarded. The retained reads were aligned to the *O. niloticus* reference genome assembly [Genbank accession number GCA_001858045.2 (Conte et al., 2017)] using bowtie2 (Langmead and Salzberg, 2012). Stacks v2 (Catchen et al., 2011; Rochette et al., 2019) was used to identify and extract single nucleotide polymorphisms (SNPs) using gstacks (settings: *–var-alpha 0.001 –gt-alpha 0.001 –min-mapq 40*). Stacks v2 primarily identified ddRAD loci corresponding to restriction enzyme cutting sites using a sliding window strategy (1 Kbp in length) in the sets of aligned reads on each sample iteratively. Upon data acquisition from all samples on each tested locus, the window was advanced to the next read beyond the previous window bound (Rochette et al., 2019). SNP calling was performed using a Bayesian genotype caller (BGC) allowing a per-nucleotide sequencing error (Maruki and Lynch, 2017). During variant calling, for numerical stabilization reasons a sequencing error under the assumption of polymorphism of at least 0.1 was assumed and the obtained genotype likelihoods were rescaled in order to be greater or equal to 1 (Rochette et al., 2019). Only one single SNP from each individual ddRAD locus was considered for downstream analysis in order to minimize the possibility of genotypic errors and reduce computational time. SNPs with a minor allele frequency (MAF) < 0.05 within a population were discarded. Finally, only SNPs that were detected in at least 75% of the samples in each population were retained for downstream analysis. The aligned reads in the format of bam files were deposited in the National Centre for Biotechnology Information (NCBI) repository under project ID PRJNA518067. The accession numbers of samples analyzed in this study are given in **File S1**.

Genetic Similarity and Relationship Among Populations

Mean observed (Ho) and expected (He) heterozygosity and average individual inbreeding coefficients (Fis) were estimated using Stacks v2 (Rochette et al., 2019). The R package StAMPP (Pembleton et al., 2013) was used to perform an Analysis of Molecular Variance (AMOVA) using 100 permutations. Additionally, pairwise F_{st} values were obtained using the stamppF$_{st}$ function according to Cockerham and Weir (1984). Furthermore, confidence intervals and p-values of the pairwise F_{st} values testing for significant deviations from zero were estimated using 1,000 bootstraps. Principal component analysis (PCA) was carried out using the R package ADEGENET version 2.1.1 (Jombart et al., 2018).

Genetic Structure and Admixture

In this study, discriminant analysis of principal components (DAPC) and Bayesian-model-based approaches were used to infer the genetic structure of *O. niloticus* samples derived from 7 populations in Tanzania. Population structure and potential admixture between the different populations was evaluated using Bayesian clustering approaches implemented in the program Structure v2.3.4 (Pritchard et al., 2000). The number of clusters tested (K) ranged from 1 to 9. Markov chain Monte Carlo of 200,000 iterations with a burn-in period of 100,000 were carried for each K-value. The delta-K method based on the criteria proposed by Evanno et al. (2005) and the obtained posterior probability values (Pritchard et al., 2000) were used to determine the optimal number of clusters. Structure results were interpreted using Structure Harvester (Earl, 2012) and CLUMPAK (Kopelman et al., 2015) for identifying the most probable number of clusters. Population structure was further confirmed using DAPC as demonstrated by Jombart et al., (2010). DAPC transformed the data using a prior PCA step and subsequently applied a discriminant analysis step (Jombart and Collins, 2015). The Bayesian Information Criterion (BIC) was

used for selecting the optimal number of clusters (K) based on the elbow method (Jombart et al., 2010).

Population Assignment and Diagnostic Single Nucleotide Polymorphisms

A four-fold cross-validation scheme was applied using the R package ADEGENET version 2.1.1 (Jombart et al., 2018) in order to test the efficiency of the SNP dataset for correctly identifying the population of origin of putatively unknown tilapia samples. The population of origin of 25% of individual fish from each genotyped population (five animals per population) was masked and was used as a test dataset. Predictions regarding population of origin on the aforementioned test set were performed using information obtained through DAPC (*predict.dapc)* on the remaining training data set. The entire procedure was repeated 10 times in order to minimize potential bias due to sample allocation in the training/test datasets. Furthermore, DAPC carried out on the entire dataset was used to identify SNPs with highest population discriminatory value.

RESULTS

Double-Digest Restriction Site-Associated DNA Sequencing and Single Nucleotide Polymorphism Identification

A total of 169 million raw sequence reads (150 bp paired-end) were obtained. Approximately 140 million reads from 139 individual fish (one fish was removed due to sequencing failure) passed the aforementioned quality control (QC) filters. Alignment of these filtered reads to the Nile tilapia reference genome (Conte et al., 2017) resulted in a mapping rate of 94–97%. In total, 31,602 putative ddRAD loci corresponding to the restriction enzymes cutting sites were identified out of which 6,779 loci were polymorphic. Derived loci had a mean sequence coverage of 120X (SD = 60X). 3,821 polymorphic sites were removed due to missing values (>25%). In addition, 778 polymorphic loci were discarded due to low MAF values (<0.05). A total of 2,180 SNPs with a MAF above 0.05 across all samples (**Figure 2**) and found in more than 75% of the genotyped fish on each population were retained for downstream analysis. The mean MAF within populations ranged from 0.07 (Kunduchi) to 0.17 (TAFIRI).

Genetic Similarity and Relationship Among Populations

The overall mean expected heterozygosity within populations was 0.132, while the observed heterozygosity was 0.081 (**Table 1**). Expected heterozygosity ranged from 0.057 in the FETA population to 0.214 in the Kunduchi population, while observed heterozygosity ranged from 0.057 in FETA to 0.113 in Ruhila (**Table 1**). Inbreeding coefficient (Fis) values ranged from low values in Lake Victoria (0.005), FETA (0.006), and Igunga (0.009) to relatively high values in Karanga (0.265), Ruhila (0.275), and Kunduchi (0.557).

Principal component analysis (PCA) was used to visualize individual relationships within and between populations. The first and second principal components accounted for 62% and 14% of the total variation, respectively. Individual fish from FETA, Lake Victoria, Igunga and most of the individual fish

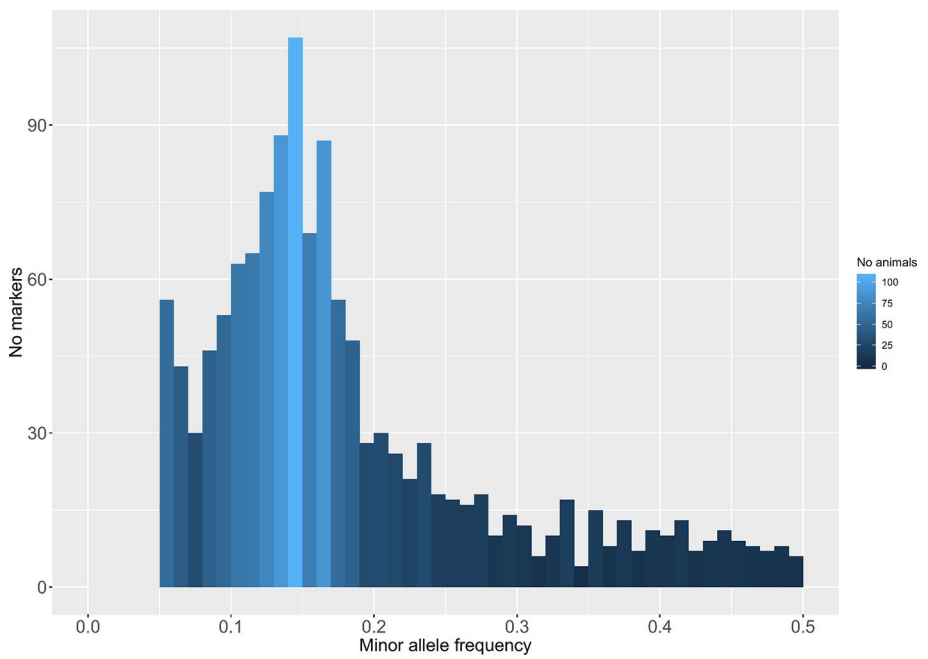


FIGURE 2 | Distribution of minor allele frequencies of double-digest restriction site-associated DNA (ddRAD)–derived single nucleotide polymorphisms (SNPs) in seven populations of Nile tilapia.

TABLE 1 | Summary of diversity parameters for the seven Nile tilapia populations.

Population	He	Ho	Fis
Ruhila	0.212	0.113	0.275
Karanga	0.213	0.104	0.265
TAFIRI	0.1	0.096	0.021
Kunduchi	0.214	0.067	0.557
Igunga	0.065	0.064	0.009
Lake Victoria	0.061	0.061	0.005
FETA	0.057	0.057	0.006

TABLE 2 | Pairwise F_{st} with 95% confidence intervals (CI) among the seven population: TAFIRI, Ruhila, FETA, Lake Victoria, Karanga, Igunga and Kunduchi.

Population 1	Population 2	Lower CI	Upper CI	F_{st}
TAFIRI	Ruhila	0.17890	0.20797	0.19322
TAFIRI	FETA	0.42571	0.47996	0.45256
TAFIRI	Victoria	0.40533	0.45925	0.43243
TAFIRI	Karanga	0.42979	0.45994	0.44498
TAFIRI	Igunga	0.39132	0.44652	0.41922
TAFIRI	Kunduchi	0.24015	0.27814	0.25967
Ruhila	FETA	0.24998	0.27900	0.26439
Ruhila	Victoria	0.23112	0.25824	0.24545
Ruhila	Karanga	0.19066	0.21102	0.20097
Ruhila	Igunga	0.22390	0.25150	0.23791
Ruhila	Kunduchi	0.06998	0.08676	0.07830
FETA	Victoria	0.08070	0.10897	0.09429
FETA	Karanga	0.53479	0.56045	0.54758
FETA	Igunga	0.03443	0.05096	0.04283
FETA	Kunduchi	0.10584	0.12052	0.11242
Victoria	Karanga	0.52577	0.55252	0.53849
Victoria	Igunga	0.02878	0.04490	0.03670
Victoria	Kunduchi	0.10591	0.11914	0.11226
Karanga	Igunga	0.51982	0.54576	0.53282
Karanga	Kunduchi	0.30078	0.32315	0.31192
Igunga	Kunduchi	0.08748	0.09959	0.09326

from Kunduchi formed a group of genetically similar animals (**Figure 3**). All TAFIRI fish formed a different group and were distinct from the other populations, except for one individual. Three individual fish from Kunduchi, one from TAFIRI, seven from Ruhila, and eight from Karanga did not group with the majority of animals from the same sampling locations.

The population pairwise F_{ST} values varied from 0.037 to 0.548 (**Table 2**). Lowest F_{ST} values were between Igunga and populations from the Lake Victoria and FETA. On the other hand, the highest F_{ST} values were between Karanga and the three most geographically distant populations, FETA, Lake Victoria and Igunga (F_{ST} = 0.548, 0.538, and 0.533 respectively). In addition, analysis of molecular variance (AMOVA) was used to detect within and among populations genetic variance components. AMOVA showed the highest levels of genetic variation within populations 67%, of the total variation, and 33% of variation was distributed among populations.

Population Genetic Structure

The STRUCTURE analysis suggested that K = 7 was the most probable number of separate clusters for the studied Nile tilapia populations. Further, individual fish from FETA, Lake Victoria, Igunga and most of individual fish from Kunduchi (16 animals) appeared to share the same genetic cluster, while animals from TAFIRI formed a separate isolated cluster (**Figure 4**). Samples from the Karanga and Ruhila populations provided evidence of admixture. In addition, the existence of unique genetic clusters is suggested for both the Karanga and Ruhila populations. The aforementioned population structure was further validated in the DAPC analysis (**Figure 5**).

Population Assignment and Diagnostic Single Nucleotide Polymorphisms

The identified SNP dataset was used for predicting the population of origin of putative unknown samples. An

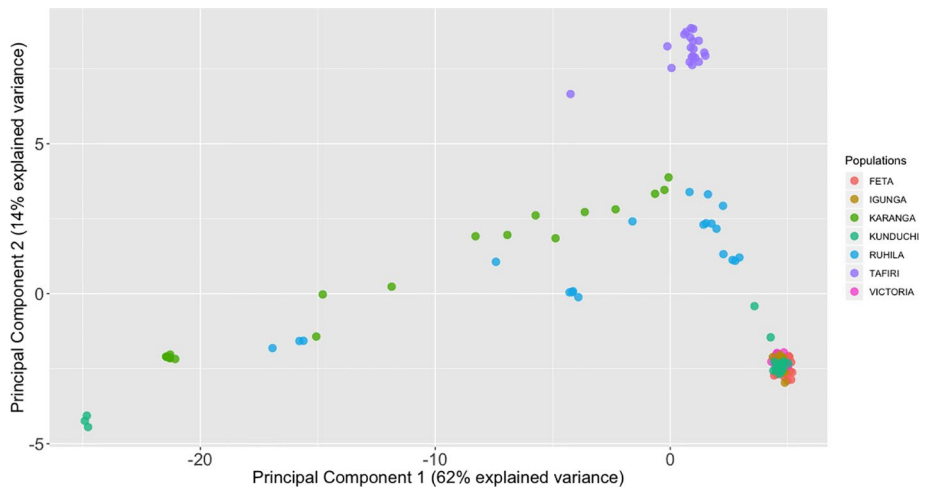


FIGURE 3 | Principal components analysis (PCA) of the population for 139 fish individual fish based on 2,180 single-nucleotide polymorphisms (SNPs). The genetic relationships among individual fish as seen when plotting the first and second principal components (PCA1 and PCA2). Each individual is represented by one dot, with its symbol color corresponding to the assigned population.

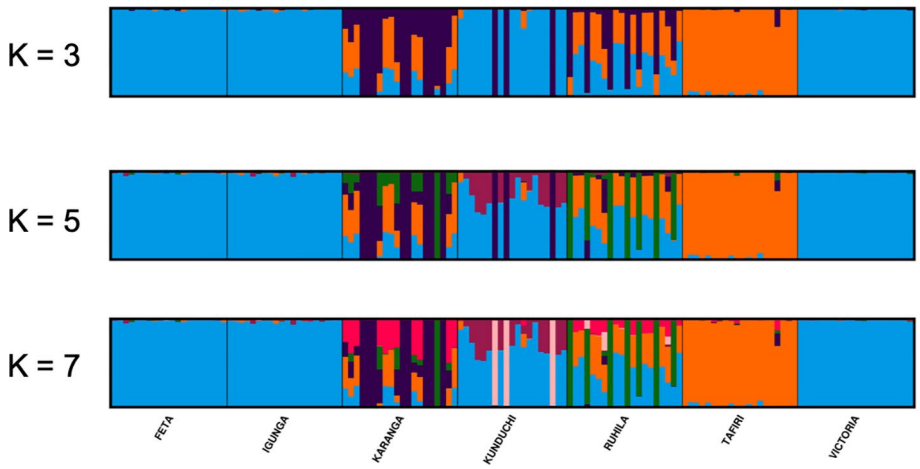


FIGURE 4 | STRUCTURE analysis bar plots for K = 3, 5, and 7 (admixture model) showing population structure of different Nile tilapia sub-populations. Each vertical stripe represents an individual. Each color represents the proportion of membership with regard to the each assigned to seven genetic clusters. Same color in different individual fish indicates that they belong to the same cluster.

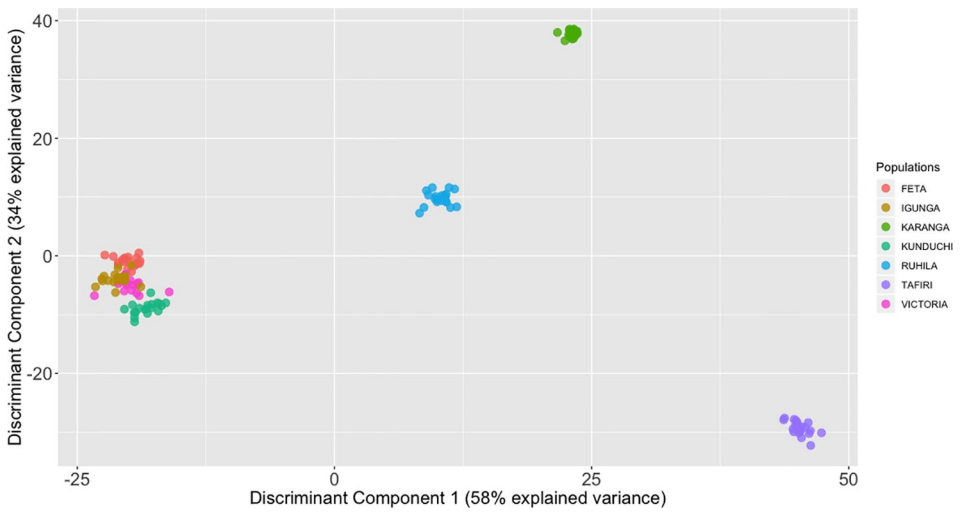


FIGURE 5 | Discriminant analysis of principal components (DAPC) analysis with the *find.clusters* for 139 individual fish of the *O. niloticus* cultured in Tanzania. The axes represent the first two linear discriminants (LD). Squares represent groups and dots represent individual fish. Numbers represent the different populations identified by DAPC analysis.

assignment rate of 77% was observed from the four-fold cross-validation analysis. The lowest correct allocation was obtained for samples from Lake Victoria, Kunduchi and Igunga (**Figure 6**). Mistakenly allocated samples were in all cases predicted as originating from either three populations (Lake Victoria, Kunduchi and Igunga). The aforementioned populations had the lowest genetic diversity values among them and formed a single cluster in the population structure analysis. In addition, DAPC analysis detected two SNPs with highest value for population identification. SNP-23095_6 and SNP-7137_40 had the highest population discriminatory value, indicating that they are the ones contributing most to cluster identification.

DISCUSSION

Understanding the patterns and extent of genetic divergence is essential both for efficient management of wild populations and for aquaculture activities. Many natural populations in Africa are under threat due to habitat destruction, overfishing and unregulated fish transfers (Eknath and Hulata, 2009). Furthermore, despite the value of Nile tilapia for the aquaculture sector in Tanzania limited research has been conducted regarding the genetic diversity of Nile tilapia populations in the country. The advent of ddRAD-seq and similar platforms have provided a cost effective and efficient technique for high resolution population genomic studies in many species (Peterson

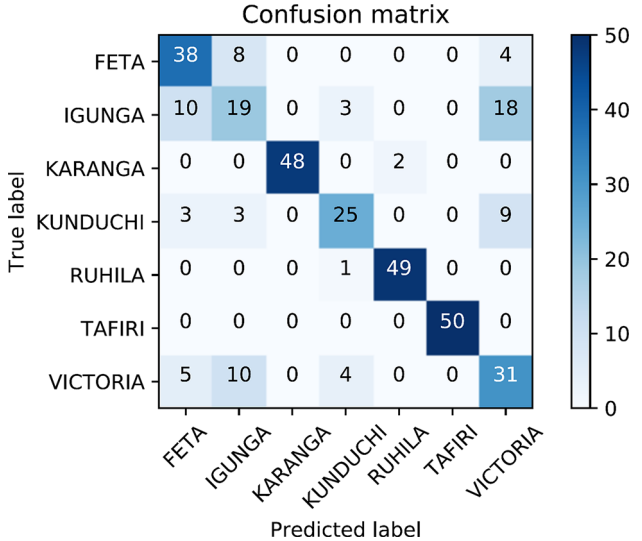


FIGURE 6 | Confusion matrix for prediction efficiency of the single-nucleotide polymorphism (SNP) dataset using cross-validation. Four-fold cross-validation was performed where five randomly chosen animals on each population were considered of unknown origin. The entire procedure was repeated 10 times in order to minimize potential bias due to sample allocation in the training/test datasets. The diagonal contains the number of correct population assignments for the overall sum of the cross-validation scheme. Off-diagonals contain the number of erroneously population allocations for each particular case.

et al., 2012; Robledo et al., 2018). In this study, 2,180 SNP markers derived from ddRAD-seq were used to assess the genetic diversity and population structure of both locally cultured and wild *Oreochromis niloticus* strains in Tanzania.

From a farming perspective, evaluation of the genetic diversity among and within tested populations is crucial in order to ensure that the most diverse animals are chosen for selective breeding practices. Since Tanzania is a hot spot for tilapias, knowledge regarding genetic diversity will also be useful in appropriate management of wild populations. In addition, genetic variation is important for a population's adaptation capacity towards changing environmental conditions (Fischer et al., 2017). Mireku et al. (2017) found higher genetic variation within populations than among populations in Nile tilapia populations from Lake Volta in Ghana. In this study AMOVA revealed the existence of higher genetic variation within populations than between populations. This could highlight that the usage of molecular markers (*e.g.* SNP data) would be of importance in future selective breeding practices as it would allow to utilize more efficiently the within population variance as opposed to traditional pedigree practices solely relying on the usage of passive integrated transponder tags. Nevertheless, as revealed by STRUCTURE analysis it should be taken into account that some populations contain unique genetic clusters not represented by "pure" populations.

Heterozygosity is a commonly used metric to compare the amount of genetic variation within different populations (Templeton and Read, 1994; Gu et al., 2014). Two different measures of heterozygosity are commonly used the observed and the expected heterozygosity. Gu et al. (2014) found that observed heterozygosity (Ho = 0.4483) in six *Oreochromis* populations in the primary rivers of Guangdong province were lower than the expected heterozygosity (He = 0.7097). On the contrary, Mireku et al. (2017) showed that observed heterozygosity (Ho = 0.526) of nine populations of *O. niloticus* in the Volta lake of Ghana was slightly higher than the expected heterozygosity (He = 0.459). In addition, Hassanien and Gilbey (2005) reported that the average of expected and observed heterozygosity were higher in *O. niloticus* populations from river Nile (He = 0.884 and Ho = 0.815) than from Delta lake populations (He = 0.846 and Ho = 0.533). In our study the overall observed heterozygosity (Ho = Ho = 0.081) was lower than the expected heterozygosity (He = 0.132) for most tested populations. Even though our study used SNP markers opposed to the aforementioned studies where microsatellites were primarily used the heterozygosity values are low compared to ddRAD studies in other fish species ranging between 0.18 and 0.25 (Saenz-Agudelo et al., 2015). A possible explanation could be due to the low MAF in our SNP dataset. In particular, over 80% of the utilized SNPs had MAF below 0.2. In addition, our results could be partly explained due to the occurrence of non-random mating. Furthermore, the low heterozygosity levels could be explained by the Wahlund effect (Wahlund, 1928) where observed heterozygosity is reduced as populations diverge. We need also to acknowledge the potential influence of the relatively small to moderate sample size for each population (20 animals per population). Nevertheless, estimates of heterozygosity from empirical data are relatively insensitive to sample size (Allendorf and Luikart, 2007).

Populations from FETA, Lake Victoria and Igunga showed the same level of expected and observed heterozygosity suggesting that random mating potentially occurred (Templeton and Read, 1994). This is further supported by the low values of inbreeding coefficients (Fis) in the populations of Igunga, FETA and Lake Victoria. High positive Fis values indicate the existence of non-random mating or population subdivision. An additional explanation for the above could be also due to the existence of null alleles. Nevertheless, since the observed excess of homozygotes appears to occur on a population level rather than locus specific we would not expect the observed excess of homozygotes to be due to the existence of null alleles. The higher diversity in Kunduchi, Karanga and Ruhila populations on the other hand may be due to both the existence of non-random mating and due to a higher degree of admixture as revealed by the STRUCTURE analysis.

Genetic differentiation among populations is further affected by migration, mutation, drift, habitat heterogeneity and selection (Holsinger and Weir, 2009). Thus the actual levels of differentiation will be a balance between the homogenizing effects of gene flow due to the former and the disruptive effects of the latter (Allendorf and Luikart, 2007). Low-moderate levels of differentiation (Fst = 0.074) have been reported between the wild Nile tilapia from Lake Volta and the improved Akosombo strains in Ghana (Mireku et al., 2017). Also low degree of differentiation (Fst = 0.0297) was found between Nile tilapia populations from rivers of the Guangdong province in China. In our study genetic differentiation among FETA, Igunga and Lake Victoria populations was particularly low

(F_{ST} values: 0.043 and 0.037 respectively). The similarity among these three populations is probably due to their origin from the same region of Lake Victoria (personal communication with fish farmers). According to our records the parents of the genotyped fish from FETA and Igunga also originated from Lake Victoria. Therefore, it is likely that these populations are genetically similar to each other and share the same genetic background. Moreover, the assignment of FETA, Lake Victoria and Igunga in the same cluster according to both STUCTURE and DAPC analysis provides further support for the aforementioned hypothesis. Nevertheless, in the case of TAFIRI a different trend was observed despite originating from the same location. The high F_{ST} values between TAFIRI and other populations (FETA, Igunga, and Lake Victoria) indicate high isolation between them. Interestingly, the TAFIRI population was composed of animals being in captivity for 4–6th generations (personal communication with a fish farmer) and this could be a reason for its genetic uniqueness. Furthermore, we observed strong genetic differentiation between Karanga and the three closely related populations of FETA, Igunga, and Lake Victoria (F_{ST} = 0.548 0.538 0.533 respectively). The differences could be the result of geographical isolation which probably has acted as a barrier to gene flow between those populations, leading to the suggested genetic structure that the STRUCTURE analysis revealed. Nevertheless, gene flow is expected to have occurred among the admixed populations (Karanga, Ruhila, Kunduchi) and expected "pure" populations of Lake Victoria and TAFIRI. Since reproductive viable hybrids in tilapias are common (Wohlfarth and Hulata, 1983), the observed admixture in Karanga population could alternatively indicate that some animals could have been mistakenly described as pure Nile tilapia. Lowe et al. (2000), reported that it is particularly difficult to identify hybrids between the species based on morphology.

Multiple approaches using both multivariate analysis (PCA, DAPC) and Bayesian clustering algorithms (STRUCTURE) were used in the current study for deriving the underlying genetic structure among the sampled populations. PCA offers considerable advantages, since it can be applied in large datasets at a minimal computational cost compared to Bayesian approaches. In general terms, PCA aims to summarize the total variation between individuals in a reduced dimension. Nevertheless, the above approach does not necessarily provide optimal resolution for distinguishing between different groups. As such, approaches like DAPC have been shown to be particularly advantageous, since they retain the computational advantages of PCA, while at the same time offer higher resolution for detecting groups of individuals with common genetic background (Jombart et al., 2010). Animals from Kunduchi, Lake Victoria, FETA and Igunga clustered together. In contrast, fish from the TAFIRI population showed greater genetic differentiation appearing separated from the other populations. Interestingly, animals from TAFIRI did not group together with FETA, Igunga and Lake Victoria despite the fact that all the populations were sampled from the same region. Differences in allele frequencies between TAFIRI and other populations might be due to the use of relatively few founder stocks and possibly unforeseen reproductive bottlenecks. Other reasons could be due to founder effects and genetic drift because of small number of parents used for breeding.

Admixture analysis further supported that FETA, Lake Victoria, and Igunga together with animals from Kunduchi shared similar genetic background. On the other hand, high admixture levels were inferred in the Karanga, Ruhila, and Kunduchi populations. In the Ruhila population admixture with the population from Lake Victoria and TAFIRI was suggested. Moreover, a similar result was obtained for the Karanga population, while in the case of Kunduchi admixed fish shared genome variation with populations of FETA, Lake Victoria, and Igunga. The speculated uncontrolled movement of fish between different locations in- and outside Tanzania, maybe from Kenya or Thailand, could be an explanation for the suggested population admixture. Nevertheless, it needs to be stressed that both Ruhila and Kunduchi appear to contain animals of a distinct genetic background.

It should be stressed that the Ruhila aquaculture development center located in the southern part of Tanzania, stocked fish from Kingolwira aquaculture center in Morogoro in 2011. The Kingolwira aquaculture center obtained their broodstock from Lake Victoria. Native species to Lake Victoria are *O. esculentus* and *O. variabilis* while *O. leucostictus* and *O. niloticus* were introduced in the lake in 1950s (Bradbeer et al., 2018). Furthermore, Shechonge et al. (2019) found evidence of introduced *Oreochromis leucostictus* males from Ruhila government pond in Songea and also reported that fish farmers misidentified *O. leucostictus* as *O. niloticus*. Additionally, in the case of Karanga population native *Oreochromis* species found in Pangani basin including Lake Jipe are *O. jipe, O. pangani* and introduced *O. niloticus* and *O. esculentus* (Shechonge et al., 2019). As such species available at Karanga station are *O. pangani*, *O. niloticus, O. jipe* and probably hybrids of three species. This could explain the high admixture level in Karanga populations compared to other populations. Overall, the high suggested admixture level for Ruhila and Karanga populations could be due to potential mislabeled samples that were wrongly classified as Nile tilapia.

The current study attempted to investigate the efficiency of the SNP dataset for population discrimination purposes of potentially unknown origin samples using a cross-validation scheme. The ability to predict the population of origin is most valuable both for fish farming practices and for conservation purposes of wild populations. Separating the dataset in a training and a validation set was applied in order to minimize overfitting, a commonly encountered situation especially in models with a considerable larger size of predictors (SNP data) than samples (genotyped fish). Model overfitting in our case could mistakenly lead to the conclusion that the SNP dataset would be highly efficient in deciphering the most probable population of origin of unknown samples. Overall 77% of tested individual fish were correctly allocated to population of origin using the SNP data. Most of the erroneous assignments originated from the three closely related populations for which our information suggests that all three originate from Lake Victoria. Further, a low number of correctly assigned individual fish were obtained in the Kunduchi population. As suggested both by STRUCTURE and DAPC high level of admixture is suggested for the Kunduchi population. Taking the above into account successful assignment to population of origin exceeded 92%. Nevertheless, it needs to be acknowledged that for the conducted analysis to be most

efficient the population information of the training dataset should be highly accurate. The expected unregulated transferring of fish in Tanzania coupled with the inherent difficulty of species discrimination among tilapias using phenotypic criteria and the most common hybridization between tilapia species resulting to reproductive viable offspring could suggest that potentially mislabeled samples have been included.

Overall, the obtained results from our study indicate that the genetic diversity and structure of Nile tilapia populations cultured in Tanzania can be explained by their life history and geographical distribution. The results also revealed greater genetic diversity within than among populations. The close clustering of Igunga, FETA and Lake Victoria populations and distinct separation of TAFIRI, suggests that these could be pure populations without admixture. The above should be taken into consideration in future wild populations conservation practices. Moreover, the gained information regarding population structure among the tested tilapia populations is important for characterizing genetic similarities and relationships of cultured lines in Tanzania. Understanding how genetic variation is distributed within and among populations will facilitate the formation of a base population and will allow breeders to design crossings between the aforementioned populations in order to maximize the genetic diversity for selective breeding purposes. Therefore, the results from this study could be used as a guide for future breeding programs and genetic improvement of local Nile tilapia in Tanzania, which may ultimately form an exemplar for the development of local tilapia species and breeds for aquaculture in African countries. Finally, using SNP data to infer the population of origin is of great importance not only for estimating genetic diversity but also in wild population conservation practices. There are unique tilapia species in Tanzania that must be protected and preserved. In addition, the SNP dataset developed can also be valuable for traceability purposes especially with regards to wild populations inhabiting nature protected reservoirs.

DATA AVAILABILITY STATEMENT

The aligned reads in the format of bam files were deposited in the National Centre for Biotechnology Information (NCBI) repository under project ID PRJNA518067. The accession numbers of samples analyzed in this study are given in **File S1**.

AUTHOR CONTRIBUTIONS

RK and FP carried out DNA extraction. CP and RH performed ddRAD library preparation and sequencing. DK, MM, and RK framed the study and contributed to designing the experiments. MM and AM provided valuable suggestions to the manuscript. RK and CP performed the statistical and genetic analyses. RK wrote the manuscript. DK and CP revised the manuscript. All authors approved the final draft of the manuscript.

ACKNOWLEDGMENTS

We would like to acknowledge financial support from the Swedish International Development Agency (Sida) and BBSRC Institute Strategic Program Grants (BBS/E/D/20002172 and BBS/E/D/30002275) from Roslin Institute (University of Edinburgh). Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1). The work was also partly supported by the Western Indian Ocean Marine Science Association (WIOMSA), under MARG II Grant for data analysis and manuscript writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01269/full#supplementary-material

REFERENCES

Allendorf, F. W., and Luikart, G. (2007). *Conservation and the genetics of populations* (Oxford UK: Blackwell Publishing).

Angienda, P. O., Lee, H. J., Elmer, K. R., Abila, R., Waindi, E. N., and Meyer, A. (2011). Genetic structure and gene flow in an endangered native tilapia fish (Oreochromis esculentus) compared to invasive Nile tilapia (*Oreochromis niloticus*) in Yala swamp, East Africa. *Conserv. Genet.* 12, 243–255. doi: 10.1007/s10592-010-0136-2

Antoniou, A., Kasapidis, P., Kotoulas, G., Mylonas, C. C., and Magoulas, A. (2017). Genetic diversity of Atlantic Bluefin tuna in the Mediterranean Sea: insights from genome-wide SNPs and microsatellites. *J. Biol. Res. (Thessalon)* 24, 3. doi: 10.1186/s40709-017-0062-2

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3, e3376. doi: 10.1371/journal.pone.0003376

Basiita, R. K., Zenger, K. R., Mwanja, M. T., and Jerry, D. R. (2018). Gene flow and genetic structure in Nile perch, *Lates niloticus*, from African freshwater rivers and lakes. *PloS One* 13, e0200001. doi: 10.1371/journal.pone.0200001

Bezault, E., Balaresque, P., Toguyeni, A., Fermon, Y., Araki, H., Baroiller, J. F., et al. (2011). Spatial and temporal variation in population genetic structure of wild Nile tilapia (*Oreochromis niloticus*) across Africa. *BMC Genet.* 12, 102. doi: 10.1186/1471-2156-12-102

Bhassu, S., Yusoff, K., Panandam, J. M., Embong, W. K., Oyyan, S., and Tan, S. G. (2004). The genetic structure of Oreochromis spp. (Tilapia) populations in Malaysia as revealed by microsatellite DNA analysis. *Biochem. Genet.* 42, 217–229. doi: 10.1023/b:bigi.0000034426.31105.da

Bolivar, R. B., and Newkirk, G. F. (2000). "Response to selection for body weight of Nile tilapia (Oreochromis niloticus) in different culture environments", in *Proceedings from the Fifth International Symposium on Tilapia Aquaculture*. K. Fitzsimmons, J. C. Filho. (American Tilapia Association and ICLARM: Rio de Janeiro, Brazil), 12–23.

Bradbeer, S. J., Harrington, J., Watson, H., Warraich, A., Shechonge, A., Smith, A., et al. (2018). Limited hybridization between introduced and critically endangered indigenous tilapia fishes in northern Tanzania. *Hydrobiologia* 832, 257–268. doi: 10.1007/s10750-018-3572-5.

Brown, J. K., Taggart, J. B., Bekaert, M., Wehner, S., Palaiokostas, C., Setiawan, A. N., et al. (2016). Mapping the sex determination locus in the hāpuku (*Polyprion oxygeneios*) using ddRAD sequencing. *BMC Genomics* 17, 448. doi: 10.1186/s12864-016-2773-4

Candy, J. R., Campbell, N. R., Grinnell, M. H., Beacham, T. D., Larson, W. A., and Narum, S. R. (2015). Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus, Osmeridae*), an anadromous Pacific smelt. *Mol. Ecol. Resour.* 15, 1421–1434. doi: 10.1111/1755-0998.12400

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: building and genotyping loci *de novo* from short-read sequences. *Genes Genom. Genet.* 1, 171–182. doi: 10.1534/g3.111.000240

Cockerham, C. C., and Weir, B. S. (1984). Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* 40, 157–164. doi: 10.2307/2530754

Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J., and Kocher, T. D. (2017). A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics* 18, 341. doi: 10.1186/s12864-017-3723-5

Earl, D. A. (2012). Structure harvester: a website and program for visualizing structure output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Eknath, A. E., and Hulata, G. (2009). Use and exchange of genetic resources of Nile tilapia (*Oreochromis niloticus*). *Rev. Aquacult.* 1, 197–213. doi: 10.1111/j.1753-5131.2009.01017.x

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individual fish using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Fernández, J., Toro, M. Á., Sonesson, A. K., and Villanueva, B. (2014). Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. *Front. Genet.* 5, 414. doi: 10.3389/fgene.2014.00414

Fernandes, I. M., Bastos, Y. F., Barreto, D. S., Lourenço, L. S., and Penha, J. M. (2017). The efficacy of clove oil as an anaesthetic and in euthanasia procedure for small-sized tropical fishes. *Braz. J. Biol.* 77, 444–450. doi: 10.1590/1519-6984.15015

Fischer, M. C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K. K., et al. (2017). Estimating genomic diversity and population differentiation–an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. *BMC Genomics* 18, 69. doi: 10.1186/s12864-016-3459-7

García-Ballesteros, S., Gutiérrez, J. P., Varona, L., and Fernández, J. (2017). The influence of natural selection in breeding programs: a simulation study. *Livest. Sci.* 204, 98–103. doi: 10.1016/j.livsci.2017.08.017

Genner, M. J., Turner, G. F., Ngatunga, B. P. (2018). A guide to the tilapia fishes of Tanzania. Available online at: https://martingenner.weebly.com/uploads/1/6/2/5/16250078/tanzania_tilapia_guide_edition1_2018.pdf. (Accessed December 20, 2018)

Gu, D. E., Mu, X. D., Song, H. M., Luo, D., Xu, M., Luo, J. R., et al. (2014). Genetic diversity of invasive Oreochromis spp. (tilapia) populations in Guangdong province of China using microsatellite markers. *Biochem. Syst. Ecol.* 55, 198–204. doi: 10.1016/j.bse.2014.03.035

Hassanien, H. A., and Gilbey, J. (2005). Genetic diversity and differentiation of Nile tilapia (*Oreochromis niloticus*) revealed by DNA microsatellites. *Aquacult. Res.* 36, 1450–1457. doi: 10.1111/j.1365-2109.2005.01368.x

Hassanien, H. A., Elnady, M., Obeida, A., and Itriby, H. (2004). Genetic diversity of Nile tilapia populations revealed by randomly amplified polymorphic DNA (RAPD). *Aquacult. Res.* 35, 587–593. doi: 10.1111/j.1365-2109.2004.01057.x

Holsinger, K. E., and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat. Rev. Genet.* 10, 639–650. doi: 10.1038/nrg2611

Hosoya, S., Kikuchi, K., Nagashima, H., Onodera, J., Sugimoto, K., Satoh, K., et al. (2018). Assessment of genetic diversity in Coho salmon (Oncorhynchus kisutch) populations with no family records using ddRAD-seq. *BMC Res. Notes* 11, 548. doi: 10.1186/s13104-018-3663-4

Jombart, T., and Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. *Imp. Coll. London-MRC. Cent. Outbreak Anal. Model.* Available at: http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf (Accessed October 30, 2018)

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi: 10.1186/1471-2156-11-94

Jombart T., Kamvar Z. N., Collins C., Lustrik R., Beugin M. P., Knaus B. J., et al. (2018). Adegenet: Exploratory Analysis of Genetic and Genomic Data. Available online at: https://cran.r-project.org/web/packages/adegenet/index.html. (Accessed November 15, 2018).

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

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* 15, 1179–1191. doi: 10.1111/1755-0998.12387

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

Li, H. L., Gu, X. H., Li, B. J., Chen, C. H., Lin, H. R., and Xia, J. H. (2017). Genome-wide QTL analysis identified significant associations between hypoxia tolerance and mutations in the GPR132 and ABCG4 genes in Nile tilapia. *Mar. Biotechnol.* 19, 441–453. doi: 10.1007/s10126-017-9762-8

Lowe, S., Browne, M., Boudjelas, S., and De Poorter, M. (2000). *100 of the world's worst invasive alien species: a selection from the global invasive species database.* Invasive Species Specialist Group, Auckland, 12:12. Available online at http://www.issg.org/pdf/publications/worst_100/english_100_worst.pdf.

Manousaki, T., Tsakogiannis, A., Taggart, J. B., Palaiokostas, C., Tsaparis, D., Lagnel, J., et al. (2016). Exploring a nonmodel teleost genome through rad sequencing—linkage mapping in Common Pandora, *Pagellus erythrinus* and comparative genomic analysis. *Genes Genom. Genet.* 6, 509–519. doi: 10.1534/g3.115.023432

Markert, J. A., Champlin, D. M., Gutjahr-Gobell, R., Grear, J. S., Kuhn, A., McGreevy, T. J., et al. (2010). Population genetic diversity and fitness in multiple environments. *BMC Evol. Biol.* 10, 205. doi: 10.1186/1471-2148-10-205

Maruki, T., and Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *Genes Genom. Genet.* 7, 1393–1404. doi: 10.1534/g3.117.039008

Mireku, K. K., Kassam, D., Changadeya, W., Attipoe, F. Y. K., and Adinortey, C. A. (2017). Assessment of genetic variations of Nile Tilapia (*Oreochromis niloticus L.*) in the Volta Lake of Ghana using microsatellite markers. *Afr. J. Biotechnol.* 16, 312–321. doi: 10.5897/AJB2016.15796

Oldenbroek, J. K. (2017). *Genomic management of animal genetic diversity*. (The Netherlands: Wageningen Academic Publishers).

Palaiokostas, C., Bekaert, M., Khan, M. G., Taggart, J. B., Gharbi, K., McAndrew, B. J., et al. (2015). A novel sex-determining QTL in Nile tilapia (*Oreochromis niloticus*). *BMC Genomics* 16, 171. doi: 10.1186/s12864-015-1383-x

Pembleton, L. W., Cogan, N. O., and Forster, J. W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol. Ecol. Resour.* 13, 946–952. doi: 10.1111/1755-0998.12129

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PloS One* 7, e37135. doi: 10.1371/journal.pone.0037135

Philippart, J. C., and Ruwet, J. C. (1982). "Ecology and distribution of Tilapias," in *The biology and culture of Tilapias*, vol. 7. Eds. R. S. V. Pullin, and R. H. Lowe-McConnell (Manila, Philippines: ICLARM), 15–60.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1071/bi9760317

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., and Houston, R. (2018). Applications of genotyping by sequencing in aquaculture breeding and genetics. *Rev. Aquacult.* 10, 670–682. doi: 10.1111/raq.12193

Romana-Eguia, M. R. R., Ikeda, M., Basiao, Z. U., and Taniguchi, N. (2004). Genetic diversity in farmed Asian Nile and red hybrid tilapia stocks evaluated from microsatellite and mitochondrial DNA analysis. *Aquaculture* 236, 131–150. doi: 10.1016/j.aquaculture.2004.01.026

Rochette, N. C., Rivera-Colon, A. G., Catchen, J. M. (2019). Stacks: analytical methods for paired end sequencing improve RADseq-based population genomics. *Mol. Ecol.* 28, 4737–4754. doi: 10.1111/mec.15253

Saenz-Agudelo, P., Dibattista, J. D., Piatek, M. J., Gaither, M. R., Harrison, H. B., Nanninga, G. B., et al. (2015). Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Mol. Ecol.* 24, 6241–6255. doi: 10.1111/mec.13471

Shechonge, A., Ngatunga, B. P., Bradbeer, S. J., Day, J. J., Freer, J. J., Ford, A. G., et al. (2019). Widespread colonisation of Tanzanian catchments by introduced Oreochromis tilapia fishes: the legacy from decades of deliberate introduction. 832, 235–253. doi: 10.1007/s10750-018-3597-9

Shechonge, A., Ngatunga, B. P., Tamatamah, R., Bradbeer, S. J., Harrington, J., Ford, A. G., et al. (2018). Losing cichlid fish biodiversity: genetic and morphological

homogenization of tilapia following colonization by introduced species. *Conserv. Genet.* 19, 1199–1209. doi: 10.1007/s10592-018-1088-1

Shirak, A., Cohen-Zinder, M., Barroso, R. M., Serousi, E., Ron, M., and Hulata, G. (2009). DNA barcoding of Israeli indigenous and introduced cichlids. *Isr. J. Aquac. Bamidgeh* 61, 83–88. http://hdl.handle.net/10524/19281

Sodsuk, P., and McAndrew, B. J. (1991). Molecular systematics of three tilapüne genera Tilapia, Sarotherodon and Oreochromis using allozyme data. *J. Fish. Biol.* 39, 301–308. doi: 10.1111/j.1095-8649.1991.tb05093.x

Taslima, K., Taggart, J. B., Wehner, S., McAndrew, B. J., and Penman, D. J. (2017). Suitability of DNA sampled from Nile tilapia skin mucus swabs as a template for ddRAD-based studies. *Conserv. Genet. Resour.* 9, 39–42. doi: 10.1007/s12686-016-0614-z

Templeton A.R., Read B. (1994). "Inbreeding: One word, several meanings, much confusion", in *Conservation Genetics* Eds. V. Loeschcke, S. K. Jain, J. Tomiuk (Birkhäuser, Basel) 68, 91–105. doi: 10.1007/978-3-0348-8510-2_9

Trewavas, E. (1983). *Tilapiine fishes of the genera Sarotherodon, Oreochromis and Danakilia*. (London: British Museum Natural History).

Twyford, A. D., and Ennos, R. A. (2012). Next-generation hybridization and introgression. *Heredity* 108, 179–189. doi: 10.1038/hdy.2011.68

Velo-Antón, G., Ayres, C., Rivera, A. C., Godinho, R., and Ferrand, N. (2007). Assignment tests applied to relocate individuals of unknown origin in a threatened species, the European pond turtle (*Emys orbicularis*). *Amphib. Reptil.* 28, 475–484. doi: 10.1163/156853807782152589

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus Betrachtet. *Hereditas* 11, 65–106. doi: 10.1111/j.1601-5223.1928.tb02483.x

Wessels, S., Krause, I., Floren, C., Schütz, E., Beck, J., and Knorr, C. (2017). ddRADseq reveals determinants for temperature-dependent sex reversal in Nile tilapia on LG23. *BMC Genomics* 18, 531. doi: 10.1186/s12864-017-3930-0

Wohlfarth, G. W., and Hulata, G. (1983). *Applied genetics of Tilapias. ICLARM Studies and Reviews.* (Manila, Philippines : ICLARM), 6, 26.

Wu, L., and Yang, J. (2012). Identifications of captive and wild Tilapia species existing in Hawaii by mitochondrial DNA control region sequence. *PloS One* 7, e51731. doi: 10.1371/journal.pone.0051731

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 March 2020
doi: 10.3389/fgene.2020.00203

Expression Profile Analysis of the Cell Cycle in Diploid and Tetraploid *Carassius auratus* red var.

Li Ren[1,2†], Jiahao Lu[1,2†], Yunpeng Fan[1,2], Yibo Hu[1,2], Jiaming Li[1,2], Yamei Xiao[1,2] and Shaojun Liu[1,2*]*

[1] State Key Laboratory of Developmental Biology of Freshwater Fish, Hunan Normal University, Changsha, China, [2] College of Life Sciences, Hunan Normal University, Changsha, China

OPEN ACCESS

Edited by:
Francesca Bertolini,
Technical University of Denmark, Denmark

Reviewed by:
Chuanju Dong,
Henan Normal University, China
Zhigang Shen,
Huazhong Agricultural University, China
Chenghui Wang,
Shanghai Ocean University, China
Wei Hu,
Institute of Hydrobiology (CAS), China

***Correspondence:**
Yamei Xiao
yameix@hunnu.edu.cn
Shaojun Liu
lsj@hunnu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 18 March 2019
Accepted: 21 February 2020
Published: 17 March 2020

Citation:
Ren L, Lu J, Fan Y, Hu Y, Li J, Xiao Y and Liu S (2020) Expression Profile Analysis of the Cell Cycle in Diploid and Tetraploid Carassius auratus red var. Front. Genet. 11:203. doi: 10.3389/fgene.2020.00203

Polyploidization often leads to "transcriptome shock," and is considered an important factor in evolution of species. Analysis of the cell cycle, which is associated with survival in polyploidy, has proved useful in investigating polyploidization. Here, we used mRNA sequencing to investigate global expression *in vitro* (in cultured cells) and *in vivo* (in fin and liver tissues) in both the diploid and tetraploid *Carassius auratus* red var.. Differential expression (DE) of genes in diploid (7482, 36.0%) and tetraploid (3787, 18.2%) states suggested that *in vitro* and *in vivo* conditions dramatically change mRNA expression levels. However, of the 20,771 total shared expressed genes, 18,050 (87.0%), including 17,905 (86.2%) non-differentially expressed genes (DEGs) and 145 (0.7%) DEGs between diploids and tetraploids, showed the same expression trends in both cultured cells and liver tissues. Of the DEGs, four of seven genes in the cell cycle pathway had the same expression trends (upregulated in diploids and tetraploids) in both cultured cells and liver tissues. Quantitative PCR analysis confirmed the same expression trends in the nine DEGs associated with regulation of the cell cycle. This research on common characteristics between diploids and tetraploids provides insights into the potential molecular regulatory mechanisms of polyploidization. The steady changes that occur between diploids and tetraploids *in vitro* and *in vivo* show the potential value of studying polyploidy processes using cultured cell lines, especially with respect to cell cycle regulation.

Keywords: polyploidy, *in vitro*, *in vivo*, cell cycle, mRNA expression

INTRODUCTION

Polyploidy occurs in plants, animals, and fungi (Comai, 2005; Blomme et al., 2006). It plays an important role in the evolutionary history of species by providing a large amount of genetic material, contributing to the genomic complexity, and further promoting speciation (Comai, 2005; Blomme et al., 2006; Otto, 2007). Polyploid breeding induced by artificial and natural mutagenesis is utilized to obtain cells and organisms with genome duplication, contributing to obtaining polyploid animals to achieve high genome plasticity, including allotetraploid hybrids of *Carassius auratus* red var. and *Cyprinus carpio* L. (Liu et al., 2001, 2016), polyploid channel catfish (*Ictalurus punctatus*) (Goudie et al., 1995), polyploid shellfish (Francesc et al., 2009), and autotetraploid *C. auratus* red var. × *Megalobrama amblycephala* (Qin et al., 2014).

Besides polyploid individuals, polyploidy has also been found in cells and tissues of diploid organisms, such as human muscle tissues, megakaryocytes, and hepatocytes (Parmacek and Epstein, 2009), as well as in some tissues under conditions of stress, such as aging seminal vesicle cells (Nguyen and Ravid, 2010). Additionally, polyploidy was shown to occur after administration of the drug cisplatin (Cantero et al., 2006) and the c-Jun N-terminal kinase inhibitor SP600123 (Zhou et al., 2016). Genetic instability in polyploid cells might lead to aneuploidy, thereby contributing to the formation of cancer (Storchova and Pellman, 2004). However, after self-breeding the allotetraploid progeny of *C. auratus* red var. and *C. carpio* L. for 26 generations, analysis of the chromosome number and reproductive fertility had revealed its genetic stability (Liu et al., 2001, 2016). To further study polyploid fish, the establishment of *in vitro* cell culture is necessary to analyze complex regulatory mechanisms including genome-wide additive and dominant expression in polyploid formation (Yoo et al., 2013).

Fibroblasts are the main cellular components of connective tissue, and can be easily obtained and cultured *in vitro*; they have been widely used to study the senescence of cells, cell damage, some congenital metabolic abnormalities and enzyme defects in basic medicine and clinical medicine research (Shima et al., 1980; Shima, 1988; Mahale et al., 2008; Swaminathan et al., 2016). Previously, cultured fibroblasts were obtained from the tail fin tissue of *C. auratus* red var. and their allotetraploid offspring (Huang et al., 2017). Here, we present an analysis of mRNA expression to investigate the cultured cells and tissues of diploid and tetraploid *C. auratus* red var.. We performed differential expression (DE) analysis between diploid and tetraploid samples in cultured fibroblasts and liver tissues. We also identified a number of mRNAs of differentially expressed genes (DEGs), and used quantitative (q) PCR to further confirm our findings in cultured cells and fin and liver tissues. Analysis of global expression in cultured cells and tissues should help to reveal whether *in vitro* cell lines can be used to research molecular expression and regulatory mechanisms in polyploid fish.

MATERIALS AND METHODS

Sample Preparation

All experiments were approved by the Animal Care Committee of Hunan Normal University and followed guidelines of the Administration of Affairs Concerning Animal Experimentation of China. *C. auratus* red var. was distributed in natural waters of China, and tetraploid *C. auratus* red var. × *C. carpio* L. were obtained from self-crossing of the allodiploid hybrid F_2 of *C. auratus* red var. (♀) × *C. carpio* L. (♂) (Liu et al., 2001, 2016). These individuals were bred and fed in pools under the same water temperature, dissolved oxygen content, and foraging conditions at the Engineering Research Center of Polyploid Fish Breeding and Reproduction of the State Education Ministry, China. Three individuals of each species were collected for further study.

Diploid cultured cells were obtained from the caudal fin of *C. auratus* red var., and tetraploid cultured cells were derived from the caudal fin of a tetraploid hybrid of *C. auratus* red var. (♀) × *C. carpio* L. (♂). Cells were cultured in complete growth medium composed of Dulbecco's modified Eagle's medium (Sigma) supplemented with 100 U/ml penicillin, 100 μg/ml streptomycin (Invitrogen, Carlsbad, CA, United States), 10% fetal bovine serum (Invitrogen, Carlsbad, CA, United States), 0.1% 2-mercaptoethanol (Invitrogen, Carlsbad, CA, United States), 1 mM sodium pyruvate (Invitrogen, Carlsbad, CA, United States), and 1 mM non-essential amino acids (Invitrogen, Carlsbad, CA, United States). Cells were grown in 5% (v/v) CO_2 at 28°C.

Determination of Ploidy Level

Before extracting total RNA, the ploidy level and DNA content of each sample were confirmed by flow cytometry. Diploid *C. auratus* red var. was used as a control group. Fish were anesthetized with 100 mg/L MS-222 (Sigma) before dissection. Fish tissues (∼0.2 cm^2) were quickly rinsed with 70% alcohol and washed with phosphate-buffered saline. They were then digested with 0.25% trypsin (Invitrogen, Carlsbad, CA, United States) for 15–30 min.

RNA Extraction

Total RNA was extracted from cultured cells, fin and liver tissues in accordance with a standard TRIzol protocol (Invitrogen, Carlsbad, CA, United States) after RNALater removal (Hummon et al., 2007). The purified RNA was quantified using a 2100 Bioanalyzer system (Agilent). Then, the RNA was used to obtain first-strand cDNA synthesized using AMV reverse transcriptase (Fermentas), with an oligo $(dT)_{12-18}$ primer at 42°C for 60 min and 70°C for 5 min.

Obtaining Transcriptome Data

For this study, we focused on the transcriptional regulation of *C. auratus* red var. *in vitro* and *in vivo* to investigate whether there is a difference in cell cycle regulation. Therefore, we obtained mRNA sequencing (seq) data of the liver tissue of diploid *C. auratus* red var. and tetraploid *C. auratus* red var. (♀) × *C. carpio* L. (♂) from the NCBI SRA database (SRR538839, SRR542431, SRR1535135, and SRR1536195) (Liu et al., 2016). Next, we submitted the mRNA-seq data of *in vitro* diploid *C. auratus* red var. and tetraploid *C. auratus* red var. (♀) × *C. carpio* L. cultured cells to the NCBI SRA (SRR7640867, SRR7640866, SRR7640869, and SRR7640868).

Mapping and Differential Expression Analysis

After removing read adapters and low-quality reads, quality of all clean reads of each library was assessed using the FastQC program[1]. Principal component analysis was used to examine the contribution of each gene to the separation of classes in the six liver transcriptomes based on Euclidean distances (Anders and Huber, 2010). mRNA-seq reads from each sample were mapped against the reference genome (*C. auratus* red

[1]http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, version 0.11.3

var.[2]) using TopHat with default parameters (Trapnell et al., 2012). Negative effects of background noise were removed based on the read counts (≤2) of genes in all biological replicates. To compare DE between diploid and tetraploid *C. auratus in vitro* and *in vivo*, the values of fragments per kilobase of transcript per million mapped reads (Mortazavi et al., 2008) were calculated using Cufflinks (version 2.1.0) (Trapnell et al., 2012). The false discovery rate (FDR) was used to determine the threshold *P*-value in multiple tests and analysis. Genes with FDR ≤ 0.01 and fold change (FC) > 2 were defined as the DE threshold using the DEGseq package of the R program (version 2.13) (R Foundation for Statistical Computing, Vienna, Austria) (Wang et al., 2010). DEGs were annotated using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases.

Determination of DEGs Using Quantitative RCR

Quantitative (q) PCR primers to amplify 11 cell-cycle-regulated genes (*lc3*, *smad6*, *p53*, *myc*, *gng10*, *id1*, *gng12*, *gadd45*, *jun*, *calm*, and *erg1*) were designed using conserved regions of coding sequences in the reference genome (**Supplementary Table 1**). Primers were used to detect expression with the ABI Prism 7500 Sequence Detection System (Applied Biosystems) and the following amplification conditions: 50°C for 5 min then 95°C for 10 min, followed by 36 cycles of 95°C for 15 s and 60°C for 45 s. Each test was performed three times. Relative quantification was performed and melting curve analysis was used to verify the generation of a single product at the end of the assay. Triplicates of each sample were used both for standard curve generation and during experimental assays. The relative expression of each gene was calibrated with β-actin, and relative mRNA expression data were analyzed using the $2^{-\Delta\ \Delta\ Ct}$ method (Livak and Schmittgen, 2001). The expression level of β-actin in induced tetraploid *C. auratus* red var. was estimated by the ratio of transcript abundance to the gene copy number using PCR and qPCR of DNA and RNA, respectively, extracted from cultured cells, caudal fin tissues, and liver tissues in diploid and tetraploid states. β-Actin expression was compared between diploid and tetraploid states.

RESULTS

Expression Patterns in Diploids and Tetraploids

To examine changes in the global transcriptomic profile between diploid and tetraploid *C. auratus* red var. *in vitro* and *in vivo*, 12 transcriptomes (from liver tissues and cultured cells; three individuals each from diploid and tetraploid) were obtained by paired-end sequencing. After initial adapter trimming and quality control, 535.9 million cleaned reads from the 12 libraries were obtained (**Supplementary Table 2**). Among these, 451.7 million cleaned reads were mapped against the reference genome of *C. auratus* red var.[3] using TopHat (**Supplementary Table 2**). The heatmap based on Euclidean distances clustered the diploid liver and tetraploid cultured cell samples. These results indicated significant differences in expression between liver tissues and cultured cells (**Figure 1**). The analysis of expression levels between diploids and tetraploids revealed the presence of silent transcripts based on a threshold of >10 reads for each gene (Ren et al., 2016). Four shared silent genes were detected in both diploid cultured cells and diploid liver samples, while only one shared gene was found in both of the tetraploid samples (**Supplementary Figure 1**).

DE Analysis *in vitro* and *in vivo* Using mRNA-Seq

After obtaining mapping information for all transcriptomes, we identified 20,771 shared expressed genes. To compare DE between *in vitro* and *in vivo* conditions, we performed DE analysis of diploid cultured cells and liver tissues (vs. 3 in **Figure 1A**) for all 20,771 expressed genes. A total of 3603 (17.3%) genes were found to be upregulated in diploid cultured cells, while 3879 (18.7%) were upregulated in diploid liver samples (**Supplementary Figure 2**). GO analysis of categories with the largest numbers of DEGs showed that 620 DEGs belonged to cell part (GO: 0044464) in the main category of cellular component, 1149 belonged to binding (GO: 0005488) in the main category of molecular function, and 1013 belonged to cellular process (GO: 0009987) in the main category of biological process (**Supplementary Figure 3**).

Next, we focused on differences in expression between tetraploid cultured cells and liver tissues (vs. 4 in **Figure 1A**), and identified 3787 (18.2%) DEGs (**Supplementary Figure 2**). Among these, 1258 were upregulated in cultured cells and 2529 were upregulated in the liver (**Figure 2B**). GO analysis of categories with the largest numbers of genes showed that 195 DEGs belonged to cell part (GO: 0044464) in the main category of cellular component, 557 belonged to binding (GO: 0005488) in the main category of molecular function, and 392 belonged to cellular process (GO: 0009987) in the main category of biological process (**Supplementary Figure 3**).

DE Analysis Between Diploids and Tetraploids Using mRNA-Seq

A comparison of diploids and tetraploids can provide insights into the regulatory mechanisms associated with different ploidy levels. Therefore, we focused on DE analysis between diploid and tetraploid cultured cells of 20,771 genes (vs. 1 in **Figure 1A**), and found that 19,238 (92.6%) were not DEGs while 1,532 (7.4%) were DEGs; these included 747 (3.6%) that were upregulated in diploid cultured cells and 784 (3.8%) that were upregulated in tetraploid ones (**Figure 2A**). A comparison of diploid and tetraploid liver samples (vs. 2

[2]https://www.ncbi.nlm.nih.gov/genome/?term=goldfish

[3]http://rd.biocloud.org.cn/

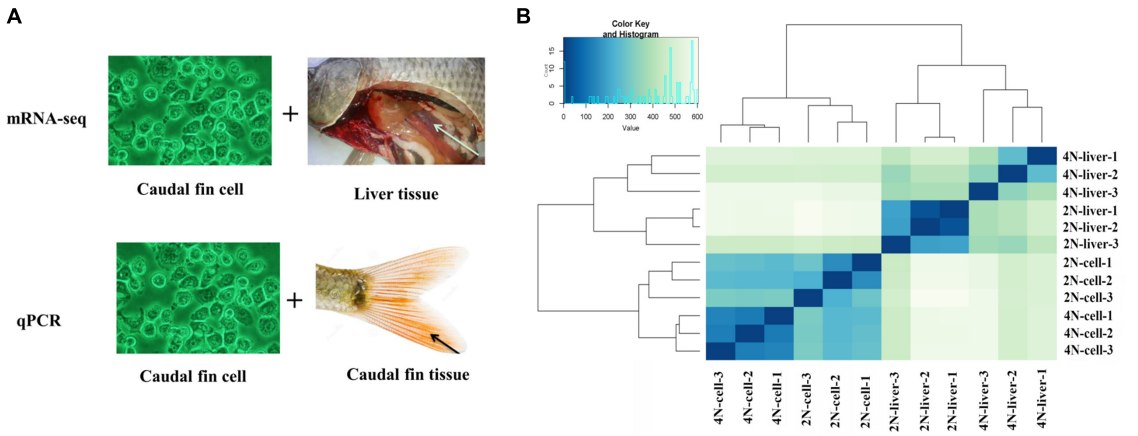


FIGURE 1 | Strategy of the expression analysis and expression cluster in all samples. **(A)** mRNA-seq and qPCR methods were used to determine the expression levels in cultured cells, caudal fin tissues, and liver tissues. The comparison of "vs. 1" and "vs. 2" was used to assess the DE of *in vivo* and *in vitro* between diploids and tetraploids. The comparison of "vs. 3" and "vs. 4" was used to assess the DE of diploids and tetraploids between *in vivo* and *in vitro*. **(B)** Overall clustering of 12 samples including diploid and tetraploid liver tissues, and diploid (2N) and tetraploid (4N) cultured cells, using normalized count data calculated by Cufflinks. The heatmap drawn from all gene count data for the reference genome depicts the relationships of all transcriptomes.

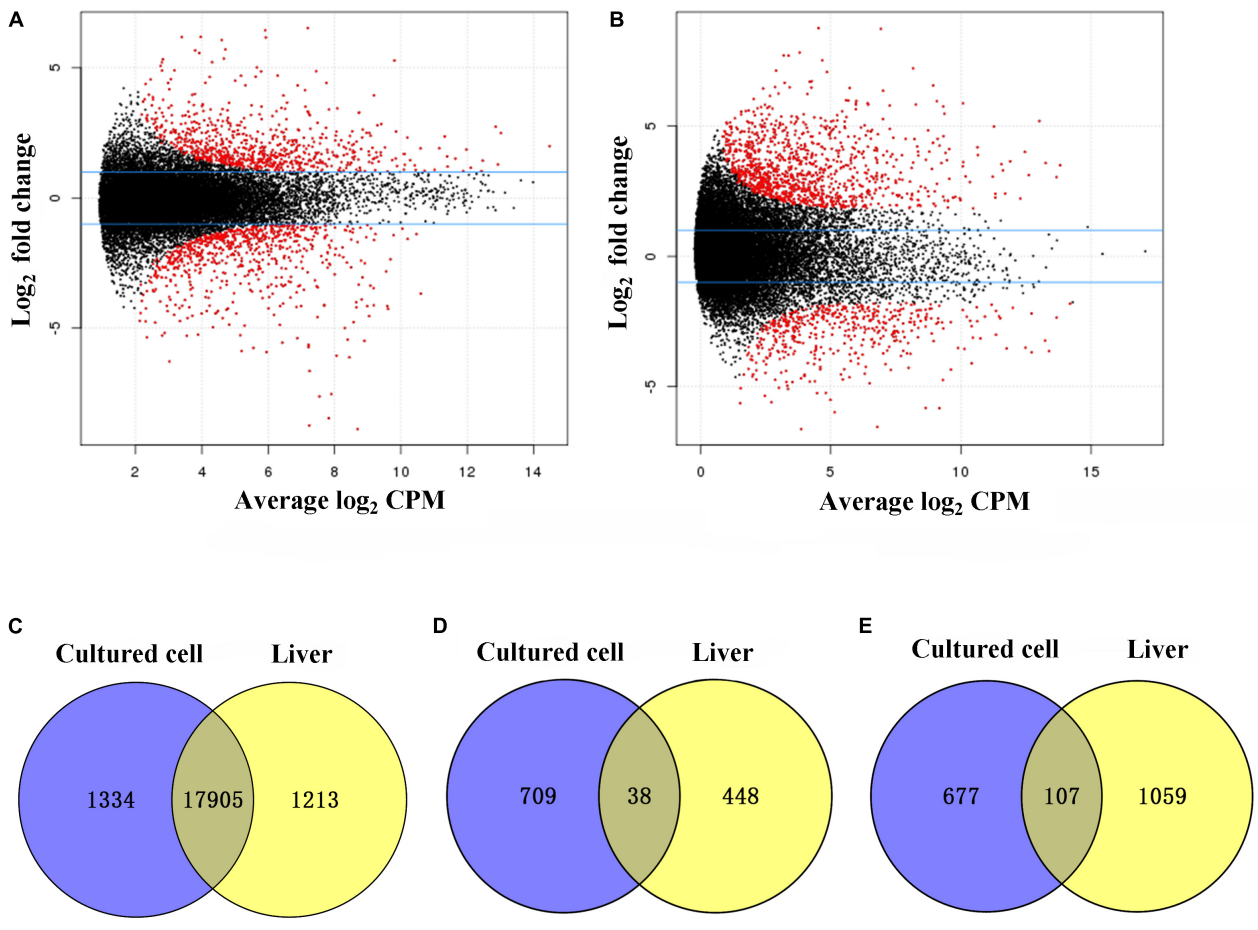


FIGURE 2 | Differentially expressed genes (DEGs) between diploid and tetraploid states in cultured cells and liver tissues. **(A)** The distribution of DEGs in cultured cells. **(B)** The distribution of DEGs in liver tissues. **(C)** Shared genes with no DE in cultured cell and liver samples. Log_2 counts per million (CPM). **(D)** Shared upregulated genes in diploid cultured cells and liver samples. **(E)** Shared upregulated genes in tetraploid cultured cells and liver samples.

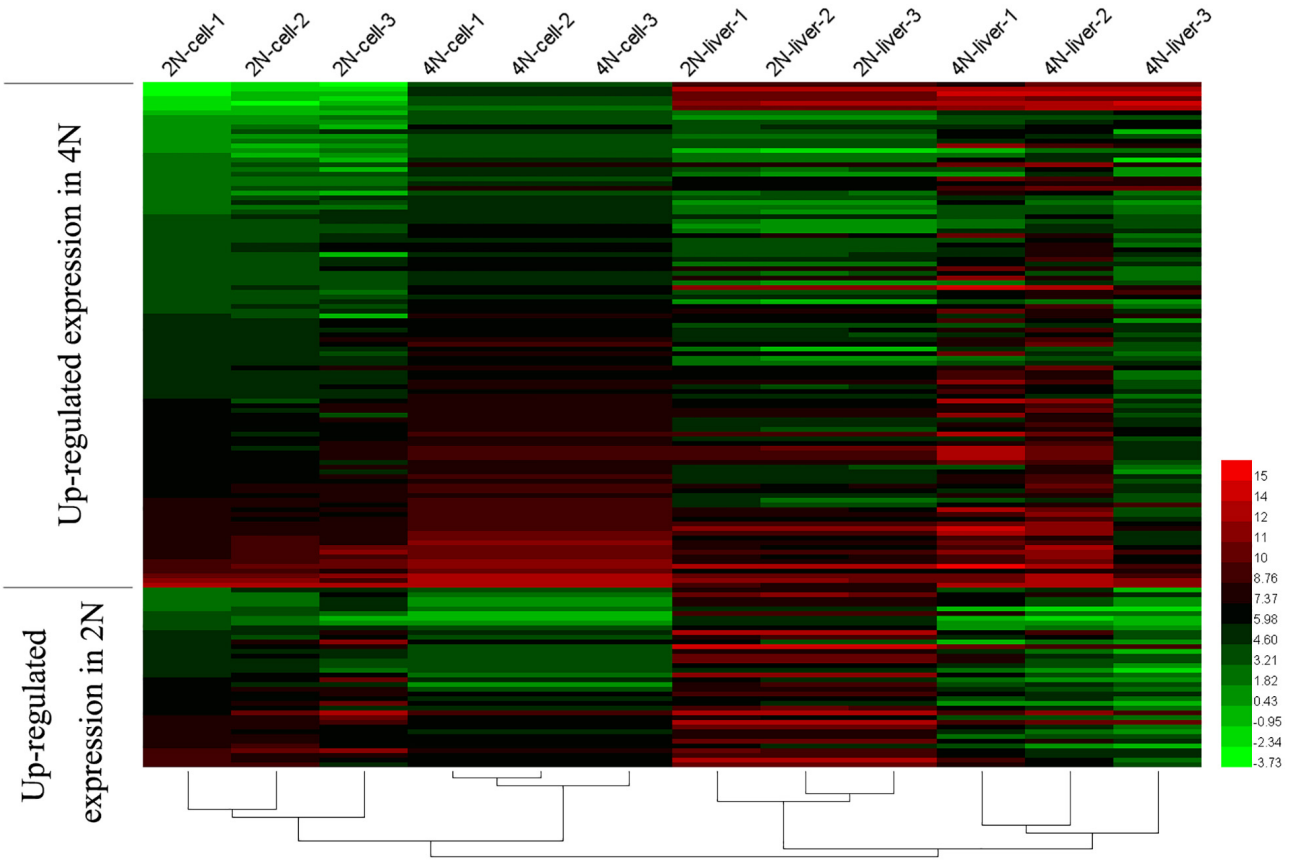


FIGURE 3 | Shared DEGs between diploids and tetraploids in cultured cells and liver tissues. The 145 shared DEGs, including 38 that were upregulated in 2N and 107 that were upregulated in 4N, were detected from the comparison between diploid and tetraploid states.

in **Figure 1A**) showed that 486 (2.3%) genes were upregulated in diploid liver tissues, while 1166 (5.6%) were upregulated in tetraploid ones (**Figure 2B**). In total, 19,238 (92.6%) and 19,048 (92.1%) genes exhibited no significant DE in cultured cells (vs. 1 in **Figure 1A**) and liver tissues (vs. 2 in **Figure 1A**), respectively. Of the 20,771 total shared expressed genes, 18,050 (87.0%), including 17,905 (86.2%) non-DEGs and 145 (0.7%) DEGs, were found to have the same expression trend in the comparisons of "vs. 1" and "vs. 2" in **Figure 1A** (**Figure 2C**). Of these 145 DEGs, 38 (0.2%) showed upregulated expression in a diploid state, while 107 (0.5%) were upregulated in a tetraploid state (**Figures 2D,E**). Additionally, the 145 shared DEGs were displayed in a heatmap, in which diploid and tetraploid liver tissue and cultured cell samples were clustered together (**Figure 3**).

DEGs Related to the Cell Cycle Pathway

To investigate changes in cell cycle regulation *in vitro* and *in vivo*, we next focused on KEGG pathways of the DEGs in our result (**Supplementary Table 3**). In comparison of diploid and tetraploid liver samples (vs. 1 and 2 in **Figure 1A**), the DEGs were shown to be mainly involved in the ribosome pathway (ko03010, 67 DEGs) and pathways associated with cancer (ko05200, 51 DEGs). Among these, 11 and 21 DEGs were associated with the cell cycle, respectively (**Figure 4**). Comparing of diploid and tetraploid cultured cells identified 15 DEGs in the cell cycle pathway. Of the seven DEGs shared between diploids and tetraploids in cultured cells and liver tissues, four showed the same expression trends as genes of the cell cycle pathway (**Figure 4**). Interestingly, three genes (*ep300a*, *myc*, and *gadd45*) exhibited the same DE trends between diploids and tetraploids, while three genes (*smad4a*, *cul1a*, and *tp53*) showed the opposite DE trends.

Expression Level Determination Using qPCR

To better investigate expression differences *in vitro* and *in vivo* (**Figure 5A**), 11 DEGs including those in cell cycle pathways were analyzed with qPCR. This was performed in cultured cells and liver tissues, as well as in fin tissue from which the cultured cells had been generated. The different conditions between cultured cells and tissues resulted in major differences in expression profiles. To better describe gene regulation in the four samples, we established expression patterns based on relative levels in cultured cells and caudal fin tissue

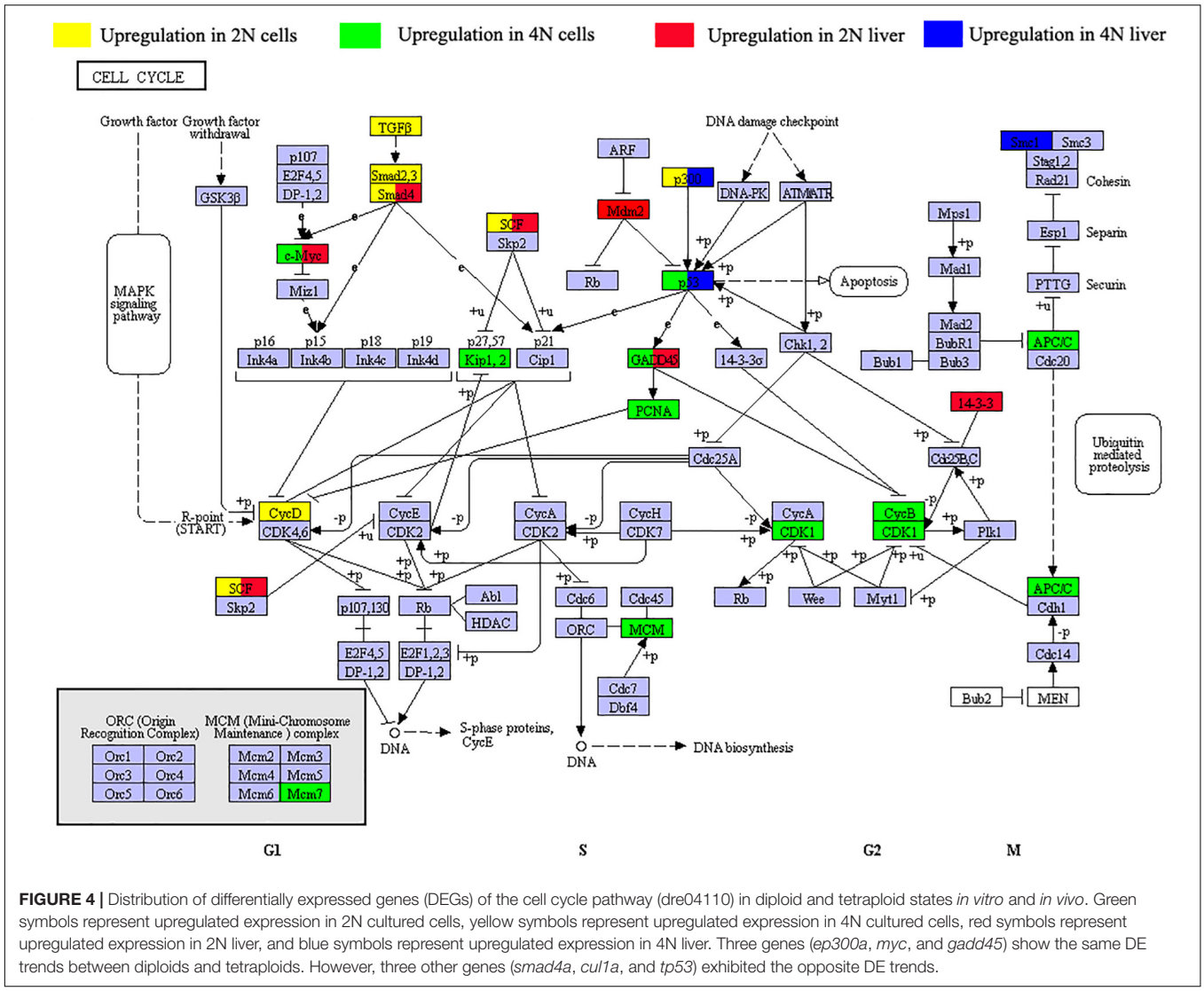


FIGURE 4 | Distribution of differentially expressed genes (DEGs) of the cell cycle pathway (dre04110) in diploid and tetraploid states *in vitro* and *in vivo*. Green symbols represent upregulated expression in 2N cultured cells, yellow symbols represent upregulated expression in 4N cultured cells, red symbols represent upregulated expression in 2N liver, and blue symbols represent upregulated expression in 4N liver. Three genes (*ep300a*, *myc*, and *gadd45*) show the same DE trends between diploids and tetraploids. However, three other genes (*smad4a*, *cul1a*, and *tp53*) exhibited the opposite DE trends.

(**Figures 5B–L**). These patterns provided a clear perspective to assess differences between diploids and tetraploids *in vitro* and *in vivo*. The same relative expression patterns between diploids and tetraploids were detected in nine genes (*smad6*, *p53*, *myc*, *id1*, *jun*, *gng10*, *gng12*, *gadd45*, and *calm*), while different relative expression patterns were detected for the two other genes (*lc3* and *erg1*) (**Figure 5**). These results of *in vitro* and *in vivo* exhibited the common trend of gene expression in cell-cycle-regulated genes accompanied with tetraploid formation.

DISCUSSION

Polyploidy were always observed in plant, but rarely in animals (Soltis et al., 2003). The formation of allotetraploid hybrids of *C. auratus* red var. and *C. carpio* L. provided an effective animal model to investigate mechanisms of polyploidy in animal (Liu et al., 2001, 2016). In comparison of diploid and tetraploid individuals, appropriate cell line were urgently needed to discover the different traits related to growth, fertility and disease resistance and various changes in molecular mechanisms for studying the potential mechanisms of these differences (Liu et al., 2001, 2009; Long et al., 2009; Ren et al., 2016). Here, we assessed the diploid and tetraploid cultured cell in gene expression level, and discussed them whether could be used to study polyploidy as comparison to *in vitro*.

Genome-wide expression profiles of polyploid culture cells and tissues in the present study provided a novel insight into the molecular mechanisms underlying the polyploidization effect *in vitro* and *in vivo*. To evaluate expression profile similarities between diploid and tetraploid states *in vitro* and *in vivo*, we performed DE analysis using mRNA-seq and qPCR. The analysis identified many DEGs between cells and liver tissues, not just in the diploid state but also in tetraploids (vs. 3 and 4 in **Figure 1A**) (**Figure 2**), indicating that marked changes in mRNA expression may be related to factors including

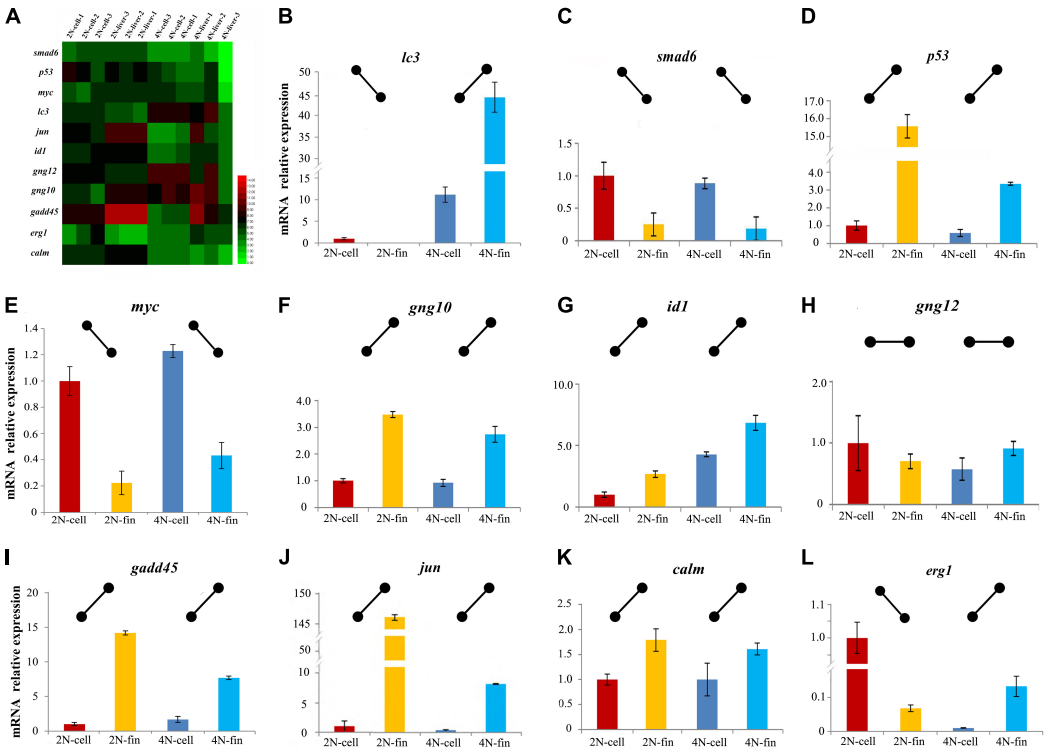


FIGURE 5 | Expression levels of 11 genes detected by mRNA-seq and qPCR. **(A)** Heatmap of the expression distribution of 11 as detected by mRNA-seq in cultured cells and liver tissues. **(B–L)** The expression levels of 11 genes detected by qPCR in cultured cells and caudal fin.

changes in the cell microenvironment and the origin of the material (Arkhipchuk and Garanko, 2005). However, in the comparison between diploid and tetraploid samples (vs. 1 and 2 in **Figure 1A**), similar expression trends, including 38 shared upregulated genes in diploids, 107 shared upregulated genes in tetraploids, and 17,905 shared genes with no DE, were found *in vitro* and *in vivo* (**Figures 2D,E**). The results preliminary suggested that the relatively stable expression trends be maintained in most genes irrespective of *in vivo* and *in vitro*.

Dramatic mRNA expression changes often occurred with hybridization and polyploidization (Leggatt and Iwama, 2003; Osborn et al., 2003; Mallet, 2007). Some DEGs distributed were observed in some aquatic organisms, including oysters (Marie et al., 2006), protogynous wrasse (Jeong et al., 2009), rice field eel (Huang et al., 2005), rainbow trout (Cleveland and Weber, 2014), and gibel carp (Sun et al., 2010; Li et al., 2014). However, gene expression of polyploid cultured cell was rarely reported. Focused on cell-cycle-regulated genes, which play an important role in cell proliferation, ontogenesis and survival (Nishihara, 1997; Ashcroft and Vousden, 2001; Boxer and Dang, 2001; Ruzinova and Benezra, 2003; Wimmer et al., 2010; Wisdom et al., 2014; Valente et al., 2015), the 11 genes had been selected and performed with expression analysis using qPCR. The same expression trends were detected in nine genes between cultured cells from fin and caudal fin tissues (**Figure 5**), further suggesting that the common trends of gene expression were in cell-cycle-regulation irrespective of *in vivo* and *in vitro*. This research focused on common characteristics between diploids and tetraploids, providing us the gene expression changes of polyploidization *in vitro* and *in vivo*. Our findings indicate that the cultured cell line of this study appears to be an appropriate platform for polyploidy research, especially into the regulation of cell proliferation and adaptive regulation, although further comparisons of diploid and tetraploid material are necessary.

DATA AVAILABILITY STATEMENT

RNA-Seq data were submitted to NCBI SRA (SRR7640867, SRR7640866, SRR7640869, and SRR7640868).

ETHICS STATEMENT

All experiments were approved by the Animal Care Committee of Hunan Normal University and followed guidelines of the Administration of Affairs Concerning Animal Experimentation of China. All samples are raised in natural ponds, dissections are performed under sodium pentobarbital anesthesia, and all efforts are made to minimize suffering. This manuscript does not involve the use of any human data or tissue.

AUTHOR CONTRIBUTIONS

LR, SL, JiamL, and YX wrote and modified the manuscript. LR, JiahL, YF, and YH provided assistance extracting the raw material, and performing the qPCR experiment and bioinformatics analyses. YX and SL contributed to the conception and design of the study. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31772902, 31730098, 31430088, U19A2040, and 31702334), the National Key Research and Development Program of China (2018YFD0901202), the Hunan Provincial Natural Science and Technology Major Project (2017NK1031), the earmarked fund for China Agriculture Research System (CARS-45), the Key Research and Development Project of Hunan Province (2016NK2128), the Scientific Research Fund of the Hunan Provincial Education Department (16C0974), the Key Research and Development Program of Hunan Province (2018NK2072), and the Research Foundation of the Education Bureau of Hunan Province (17K058).

ACKNOWLEDGMENTS

We thank Zhanzhou Yao and Rurong Zhao for technical support with experimental equipment. We also thank Sarah Williams, Ph.D., from Liwen Bianji, Edanz Group China (www.liwenbianji.cn), for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00203/full#supplementary-material

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Arkhipchuk, V. V., and Garanko, N. N. (2005). Using the nucleolar biomarker and the micronucleus test on in vivo fish fin cells. *Ecotoxicol. Environ. Saf.* 62, 42–52. doi: 10.1016/j.ecoenv.2005.01.001

Ashcroft, M., and Vousden, K. H. (2001). *Tumor Suppressor Protein.* Totowa, NJ: Humana Press. p53

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7:R43. doi: 10.1186/gb-2006-7-5-r43

Boxer, L. M., and Dang, C. V. (2001). Translocations involving c-myc and c-myc function. *Oncogene* 20:5595. doi: 10.1038/sj.onc.1204595

Cantero, G., Pastor, N., Mateos, S., Campanella, C., and Cortes, F. (2006). Cisplatin-induced endoreduplication in CHO cells: DNA damage and inhibition of topoisomerase II. *Mutat. Res.* 599, 160–166. doi: 10.1016/j.mrfmmm.2006.02.006

Cleveland, B. M., and Weber, G. M. (2014). Ploidy effects on genes regulating growth mechanisms during fasting and refeeding in juvenile rainbow trout (Oncorhynchus mykiss). *Mol. Cell. Endocrinol.* 382, 139–149. doi: 10.1016/j.mce.2013.09.024

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Francesc, P., Andy, B., Jeanclaude, F. R., Martin, F. H., Pierrick, H., and Lorenzo, C. (2009). Polyploid fish and shellfish: production, biology and applications to aquaculture for performance improvement and genetic containment. *Aquaculture* 293, 125–156. doi: 10.1016/j.aquaculture.2009.04.036

Goudie, C. A., Simco, B. A., Davis, K. B., and Liu, Q. (1995). Production of gynogenetic and polyploid catfish by pressure-induced chromosome set manipulation. *Aquaculture* 133, 185–198. doi: 10.1016/0044-8486(94)00367-W

Huang, X., Guo, Y., Shui, Y., Gao, S., Yu, H., Cheng, H., et al. (2005). Multiple alternative splicing and differential expression of dmrt1 during gonad transformation of the rice field eel. *Biol. Reprod.* 73, 1017–1024. doi: 10.1095/biolreprod.105.041871

Huang, Y., Luo, Y., Liu, J., Gui, S., Wang, M., Liu, W., et al. (2017). A light-colored region of caudal fin: a niche of melanocyte progenitors in crucian carp (Cyprinus carpio L.). *Cell Biol Int.* 41:42. doi: 10.1002/cbin.10698

Hummon, A. B., Lim, S. R., Difilippantonio, M. J., and Ried, T. (2007). Isolation and solubilization of proteins after TRIzol extraction of RNA and DNA from patient material following prolonged storage. *Biotechniques* 42, 467–470. doi: 10.2144/000112401

Jeong, H. B., Park, J. G., Park, Y. J., Takemura, A., Hur, S. P., Lee, Y. D., et al. (2009). Isolation and characterization of DMRT1 and its putative regulatory region in the protogynous wrasse. *Halichoeres tenuispinis. Gene* 438, 8–16. doi: 10.1016/j.gene.2009.03.006

Leggatt, R. A., and Iwama, G. K. (2003). Occurrence of polyploidy in the fishes. *Rev Fish Biol Fish.* 13, 237–246. doi: 10.1023/B:RFBF.0000033049.00668.fe

Li, X. Y., Li, Z., Zhang, X. J., Zhou, L., and Gui, J. F. (2014). Expression characterization of testicular DMRT1 in both Sertoli cells and spermatogenic cells of polyploid gibel carp. *Gene.* 548, 119–125. doi: 10.1016/j.gene.2014.07.031

Liu, L., Cuiping, Y., Shaojun, L., Huan, Z., Dong, L., Zhen, L., et al. (2009). Cloning and evolutionary analysis of cyclin B gene introns in cyprinids with different ploidy levels. *Prog. Nat. Sci.* 19, 1103–1108. doi: 10.1016/j.pnsc.2009.03.001

Liu, S., Liu, Y., Zhou, G., Zhang, X., Luo, C., Feng, H., et al. (2001). The formation of tetraploid stocks of red crucian carp x common carp hybrids as an effect of interspecific hybridization. *Aquaculture* 192, 171–186.

Liu, S., Luo, J., Chai, J., Ren, L., Zhou, Y., Huang, F., et al. (2016). Genomic incompatibilities in the diploid and tetraploid offspring of the goldfish x common carp cross. *Proc Natl Acad Sci U.S. A.* 113, 1327–1332. doi: 10.1073/pnas.1512955113

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Long, Y., Zhong, H., Liu, S., Tao, M., Chen, L., Xiao, J., et al. (2009). Molecular characterization and genetic analysis of Gnrh2 and Gth in different ploidy level fishes. *Prog. Nat. Sci.* 19, 1569–1579. doi: 10.1016/j.pnsc.2009.06.002

Mahale, A. M., Khan, Z. A., Igarashi, M., Nanjangud, G. J., Qiao, R. F., Yao, S., et al. (2008). Clonal selection in malignant transformation of human fibroblasts transduced with defined cellular oncogenes. *Cancer Res.* 68, 1417–1426. doi: 10.1158/0008-5472.CAN-07-3021

Mallet, J. (2007). Hybrid speciation. *Nature* 446, 279–283. doi: 10.1038/nature05706

Marie, V., Gonzalez, P., Baudrimont, M., Boutet, I., Moraga, D., Bourdineaud, J. P., et al. (2006). Metallothionein gene expression and protein levels in triploid and diploid oysters *Crassostrea gigas* after exposure to cadmium and zinc. *Environ. Toxicol. Chem.* 25, 412–418.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Nguyen, H. G., and Ravid, K. (2010). *Polyploidy: Mechanisms and Cancer Promotion in Hematopoietic and Other Cells*. New York, NY: Springer.

Nishihara, A. (1997). Smad6 inhibits signalling by the TGF-beta superfamily. *Nature* 389, 622–626. doi: 10.1038/39355

Osborn, T. C., Pires, J. C., Birchler, J. A., Auger, D. L., Chen, Z. J., Lee, H. S., et al. (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* 19:141.

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell* 131, 452–462. doi: 10.1016/j.cell.2007.10.022

Parmacek, M. S., and Epstein, J. A. (2009). Cardiomyocyte renewal. *N. Engl. J. Med.* 361, 86–88. doi: 10.1056/NEJMcibr0903347

Qin, Q., Wang, Y., Wang, J., Dai, J., Xiao, J., Hu, F., et al. (2014). The autotetraploid fish derived from hybridization of Carassius auratus red var. (female) x Megalobrama amblycephala (male). *Biol. Reprod.* 91, 93. doi: 10.1095/biolreprod.114.122283

Ren, L., Li, W., Tao, M., Qin, Q., Luo, J., Chai, J., et al. (2016). Homoeologue expression insights into the basis of growth heterosis at the intersection of ploidy and hybridity in Cyprinidae. *Sci. Rep.* 6:27040. doi: 10.1038/srep27040

Ruzinova, M. B., and Benezra, R. (2003). Id proteins in development, cell cycle and cancer. *Trends Cell Biol.* 13, 410–418.

Shima, A. (1988). Fish cell culture: establishment of two fibroblast-like cell lines (OL-17 and OL-32) from fins of the medaka, oryzias latipes. *In Vitro Cell Dev. Biol.* 24, 294–298. doi: 10.1007/BF02628830

Shima, A., Nikaido, O., Shinohara, S., and Egami, N. (1980).). Continued in vitro growth of fibroblast-like cells (RBCF-1) derived from the caudal fin of the fish, *Carassius auratus*. *Exp. Gerontol.* 15, 305–314.

Soltis, D. E., Soltis, P. S., and Tate, J. A. (2003). Advances in the study of polyploidy since Plant Speciation. *New Phytol.* 161, 173–191. doi: 10.1046/j.1469-8137.2003.00948.x

Storchova, Z., and Pellman, D. (2004). From polyploidy to aneuploidy, genome instability and cancer. *Nat. Rev. Mol. Cell Biol.* 5, 45–54. doi: 10.1038/nrm1276

Sun, M., Li, Z., and Gui, J. F. (2010). Dynamic distribution of spindlin in nucleoli, nucleoplasm and spindle from primary oocytes to mature eggs and its critical function for oocyte-to-embryo transition in gibel carp. *J. Exp. Zool. A Ecol. Genet. Physiol.* 313, 461–473. doi: 10.1002/jez.618

Swaminathan, T. R., Basheer, V. S., Gopalakrishnan, A., Sood, N., and Pradhan, P. K. (2016). A new epithelial cell line. *Cytotechnology* 68, 515–523. doi: 10.1007/s10616-014-9804-2

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Valente, G., Morani, F., Nicotra, G., Fusco, N., Peracchio, C., Titone, R., et al. (2015). Expression and clinical significance of the autophagy proteins BECLIN 1 and LC3 in Ovarian Cancer. *Biomed Res Int.* 2014:462658. doi: 10.1155/2014/462658

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612

Wimmer, P., Schreiner, S., Everett, R. D., Sirma, H., Groitl, P., and Dobner, T. (2010). SUMO modification of E1B-55K oncoprotein regulates isoform-specific binding to the tumour suppressor protein PML. *Oncogene* 29, 5511–5522. doi: 10.1038/onc.2010.284

Wisdom, R., Johnson, R. S., and Moore, C. (2014). c—Jun regulates cell cycle progression and apoptosis by distinct mechanisms. *Embo J.* . 18, 188–197. doi: 10.1093/emboj/18.1.188

Yoo, M. J., Szadkowski, E., and Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110, 171–180. doi: 10.1038/hdy.2012.94

Zhou, Y., Wang, M., Jiang, M., Peng, L., Wan, C., Liu, J., et al. (2016). Autotetraploid cell line induced by SP600125 from crucian carp and its developmental potentiality. *Sci. Rep.* 6:21814. doi: 10.1038/srep21814

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fgene.2020.00261

Conservation Genomic Analysis of the Croatian Indigenous Black Slavonian and Turopolje Pig Breeds

Boris Lukić[1], Maja Ferenčaković[2], Dragica Šalamon[2], Mato Čačić[3], Vesna Orehovački[3], Laura Iacolina[4,5], Ino Curik[2] and Vlatka Cubric-Curik[2]**

[1] Department for Animal Production and Biotechnology, Faculty of Agrobiotechnical Sciences Osijek, J.J. Strossmayer University of Osijek, Osijek, Croatia, [2] Department of Animal Science, Faculty of Agriculture, University of Zagreb, Zagreb, Croatia, [3] Ministry of Agriculture, Zagreb, Croatia, [4] Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark, [5] Department for Apiculture, Wildlife Management and Special Zoology, Faculty of Agriculture, University of Zagreb, Zagreb, Croatia

The majority of the nearly 400 existing local pig breeds are adapted to specific environments and human needs. The demand for large production quantities and the industrialized pig production have caused a rapid decline of many local pig breeds in recent decades. Black Slavonian pig and Turopolje pig, the latter highly threatened, are the two Croatian local indigenous breeds typically grown in extensive or semi-intensive systems. In order to guide a long-term breeding program to prevent the disappearance of these breeds, we analyzed their genetic diversity, inbreeding level and relationship with other local breeds across the world, as well as modern breeds and several wild populations, using high throughput genomic data obtained using the Illumina Infinium PorcineSNP60 v2 BeadChip. Multidimensional scaling analysis positioned Black Slavonian pigs close to the UK/North American breeds, while the Turopolje pig clustered within the Mediterranean breeds. Turopolje pig showed a very high inbreeding level ($F_{ROH > 4\ Mb}$ = 0.400 and $F_{ROH > 8\ Mb}$ = 0.332) that considerably exceeded the level of full-sib mating, while Black Slavonian pig showed much lower inbreeding ($F_{ROH > 4\ Mb}$ = 0.098 and $F_{ROH > 8\ Mb}$ = 0.074), indicating a planned mating strategy. In Croatian local breeds we identified several genome regions showing adaptive selection signals that were not present in commercial breeds. The results obtained in this study reflect the current genetic status and breeding management of the two Croatian indigenous local breeds. Given the small populations of both breeds, a controlled management activity has been implemented in Black Slavonian pigs since their commercial value has been recognized. In contrast, the extremely high inbreeding level observed in Turopolje pig argues for an urgent conservation plan with a long-term, diversity-oriented breeding program.

Keywords: genomics, diversity, inbreeding, local breeds, population structure

OPEN ACCESS

Edited by:
Maria Saura,
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Spain

Reviewed by:
Zhihua Jiang,
Washington State University, United States
Salvatore Mastrangelo,
University of Palermo, Italy

****Correspondence:***
Boris Lukić
blukic@fazos.hr
Vlatka Cubric-Curik
vcubric@agr.hr

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *06 May 2019*
Accepted: *05 March 2020*
Published: *31 March 2020*

Citation:
Lukić B, Ferenčaković M, Šalamon D, Čačić M, Orehovački V, Iacolina L, Curik I and Cubric-Curik V (2020) Conservation Genomic Analysis of the Croatian Indigenous Black Slavonian and Turopolje Pig Breeds. Front. Genet. 11:261. doi: 10.3389/fgene.2020.00261

INTRODUCTION

For thousands of years, pigs have been indispensable to humans as they represent an important part of our everyday diet. Pigs were domesticated about 8,500–10,500 years ago (Peters et al., 2005; Zeder, 2017) and have changed over time from their wild ancestors, especially in the last few hundred years due to the force of artificial selection. Nearly 400 local breeds have been obtained.

Pig populations continue to change genetically because of continuous gene flow between wild and domestic pigs (Iacolina et al., 2018; Frantz et al., 2019). Most local breeds are adapted to specific environments, production systems, geographical regions or human demands. However, in the last few decades, several breeds such as Large White, Duroc, Landrace, Hampshire and Pietrain (FAO, 2007) and their hybrids have spread internationally and replaced most local breeds, mainly because they are economically more efficient. This poses a problem: the conservation of local breeds is crucial for the future of animal production as they can be important sources of genetic variability (Bruford et al., 2015) and are better adapted for production in sustainable environments (Ollivier et al., 2005).

In Croatia, the two indigenous breeds, Black Slavonian (CROBS) and Turopolje (CROTS) pigs, have specific phenotypic characteristics that make them well adapted to the extensive and semi-intensive systems common in the country. The Black Slavonian breed is also known as Fajferica and was bred by the earl Karl Pfeiffer in the second half of the 19th Century in Slavonia, a "corn belt" region in Eastern Croatia. By crossing the local Mangalitza gilts with Berkshire boars, Pfeiffer created a new breed with more desirable economically important traits (feed conversion ratio, daily gain, carcass traits) than the local dominant, primitive breeds such as Mangalitza, Šiška and Bagun. Some years later, the breed was further improved by crossing the best gilts with imported USA Poland China boars. In the 1930's and 1940's, the breed was crossed again with Berkshire, and later with Large Black on several farms (Hrasnica et al., 1958). The breed was economically successful, well known for its fat and meat production and one of the most abundant (>300,000 individuals) in Yugoslavia in the 1950's (Hrasnica et al., 1958). Since then, however, the Black Slavonian pig is slowly being replaced by modern breeds such as Landrace, Large White, and Pietrain. The Black Slavonian population declined drastically at the beginning of the 1990's, during and after the war in Croatia. The first conservation program with pedigree recording started in 1996 in the founding population, consisting of the only remaining 46 sows and six boars (Uremović, 2004).

The Turopolje pig breed, for its part, is named after a small region near Zagreb (Turopolje), and has a controversial history. A publication from 1911 (Ulmansky, 1911) asserted that CROTS is a cross between local pigs and Šiška, a primitive regional breed currently extinct. A study from 1935 (Ritzoffy, 1935) claimed that Turopolje pig was most probably derived from the Slovenian Krškopolje pig at the beginning of the 19th century, while more recent work (Porter, 2002) has claimed that this breed originated from Šiška, Krškopolje and Berkshire pigs. During the second half of the 20th century, when modern breeds were intensively imported, the Turopolje breed declined severely, like many local European pig breeds. Its survival was also seriously threatened during the war in the 1990's (Druml et al., 2012).

A sustainable breeding program might prevent further erosion of the genetic adaptive capacity of both Croatian indigenous breeds and lead to more stable populations. The Black Slavonian breed is better positioned than the Turopolje breed as their carcass traits better suit today's market demands, while the Turopolje breed is a typical lard type of pig that is no longer profitable for farmers, although a recent study showed the breed has potential for some commercially relevant traits (Muñoz et al., 2018). A prerequisite for building a long-term sustainable local breeding program is detailed molecular and genomic characterization. Studies have already explored the population genetic background for the Black Slavonian breed using pedigree information (Lukić et al., 2015), microsatellite markers for both Black Slavonian and Turopolje pigs (Druml et al., 2012; Šprem et al., 2014) and selected single-nucleotide polymorphisms (SNPs) associated with morphological traits (Muñoz et al., 2018). However, a wider and more systematic approach is required to obtain more thorough understanding of their breed genetics.

Recent advances in genotyping technologies, such as SNP chips, provide affordable access to genotypic information for all major domestic animal species, enabling the estimation of the genetic diversity, population structure, genetic admixture, inbreeding level, and effective population size (Kukučková et al., 2017). In addition, SNP data can be merged and compared with the results of other studies, which is impossible with microsatellite marker analyses. Analysis of the frequencies of a large number of SNP alleles provides deep insight into genetic variability and genetic structure. For instance, the genomic inbreeding coefficient obtained from SNP analyses is more reliable than pedigree estimates (Ferenčaković et al., 2013; extensively described by Keller et al., 2011). Genetic admixture, a phenomenon that occurs when genetically divergent populations begin to interbreed (Balding et al., 2007), is conventionally identified by multivariate genetic cluster algorithms (Jombart et al., 2010).

Taking advantage of high-throughput genomic analyses, we explored the genetic structure of Black Slavonian and Turopolje pig breeds and their relationships with local and modern breeds worldwide. We also estimated the inbreeding level based on runs of homozygosity (ROH) as well as admixture level, particularly important for the Turopolje pig, which is classified as a highly endangered breed (Croatian Agricultural Agency, 2017). For each indigenous Croatian breed, we identified a set of SNPs that differentiate it from the most widespread modern commercial breeds. These results may inform future conservation management of Black Slavonian and Turopolje pigs.

MATERIALS AND METHODS

Data Collection, Quality Control, and Multidimensional Scaling

The animals in this study were selected in collaboration with the Croatian Agricultural Agency, which is the national body that manages breeding programs, and the National Gene Bank within the Ministry of Agriculture of Croatia. All procedures with animals were performed in accordance with national and European ethical protocols and directives. Animals were raised by registered breeders at more than five locations, with available information about their origin. In the case of Black Slavonian pigs, sampling of close relatives (parent-offspring, full sibs or half sibs) was avoided. In the case of Turopolje pigs, animals were sampled at random because the population was extremely small

(124 sows and 17 boars; Croatian Agricultural Agency, 2017) and contained many higher-order relatives. In this case, avoiding sampling of close relatives would lead to biased results. More detailed information describing samples in this study is provided in **Supplementary Table S1**.

A total of 16 Black Slavonian pigs (six boars and 10 sows) and 16 Turopolje pigs (four boars and 12 sows) were genotyped using Illumina PorcineSNP60 v2 Genotyping BeadChip with 64,232 SNPs (Ramos et al., 2009). DNA was isolated from hair follicles using a commercial kit (DNeasy Blood and Tissue Kits, Qiagene, Germany). Using the obtained genotypes, we analyzed only autosomal SNPs whose chromosomal position was assigned. SNPs where more than 10% of genotypes were missing and SNPs with Illumina GenCall score $\leq$ 0.7 or Illumina GenTrain score $\leq$ 0.4 (Ferenčaković et al., 2013) were excluded from the analysis. Pigs for which $>$ 5% of the genotype was missing were also excluded from further analysis. SNP positions were based on the pig genome assembly Sscrofa 10.2 (EnsEMBL db version 83). In order to compare our data with worldwide data sets, additional data (Ai et al., 2013; Burgos-Paz et al., 2013; Goedbloed et al., 2013) were downloaded from the publicly available Dryad Digital Repository (Yang et al., 2017).

We used several criteria to select breeds from public data. First, we selected breeds known to share a history with Croatian local breeds (e.g., founder breeds) or to inhabit areas close to those of Croatian breeds. Second, breeds with similar phenotypic traits such as coat color or exterior traits were selected, since such traits were among the main selection criteria during early stages of animal breeding. Genetic similarity of local breeds with wild boar is expected to be high, so we included several wild European populations. Chinese breeds were also included because of their known introgression into the international gene pool, and particularly into the commercially important breeds Landrace, Pietrain, and Duroc. This data set was then merged with our samples to produce a consensus data set containing 931 animals from 48 breeds (of which nine were wild boar populations) and 45,000 SNPs. SNP genotypes were used to calculate shared genetic coancestry between all possible pairs of individuals of all breeds in the analysis in terms of pairwise proportions of identical-by-state alleles using R software version 3.6.1 (R Core Team, 2019). The obtained matrix was transformed to a distant matrix, on which classical multidimensional scaling and principal component analysis were performed. This analysis showed that Chinese breeds, USA Feral Pig, Argentina Semi Feral Pig, Brazil Monteiro Pig, Guatemala Creole Pig, Peru Creole Pig, USA Guinea hog, USA Mulefoot, and Duroc form distant clusters (**Figure 1**). To provide better resolution and more precise characterization of the Black Slavonian and Turopolje pigs, breeds present in distant clusters were removed from subsequent analyses.

The final data set consisted of 556 animals sampled from 30 breeds, including six wild boar populations. Landrace and Pietrain breeds were represented by two different populations to provide additional controls. The following breeds were used in the analyses: Black Slavonian – CROBS (n = 16), Croatian Wild Boar – CROWB (16), Czech Prestice – TRPR (15), German Angler Sattleschwein – DEAS (n = 10), Hungarian Mangalitza – HUMA (n = 20), Iberian Wild Boar – IBWB (n = 17), Italian Calabrese – ITCA (n = 15), Italian Casertana – ITCT (n = 14), Italian Cinta Senese – ITCS (n = 13), Italian Nera Siciliana – ITNS (n = 15), Italian Sardinian Wild Boar – ITWB2 (n = 20), Italian Wild Boar – ITWB1 (n = 19), Landrace population 1 – LDR1 (n = 20), Landrace population 2 – LDR2 (n = 15), NW European Wild Boar – NEWB (n = 20), Pietrain population 1 – PIT1 (n = 20), Pietrain population 2 – PIT2 (n = 20), Polish Pulawska Spot – PLPS (n = 15), Portuguese Bisaro – PTBI (n = 14), South Balkan Wild Boar – SBWB (n = 20), Spanish Chato Murciano – ESCM (n = 20), Spanish Iberian – ESIB (n = 20), Turopolje – CROTS (n = 16), UK Berkshire – UKBK (n = 20), UK British Saddleback – UKBS (n = 20), UK Gloucester Old Spot – UKGO (n = 20), UK Hampshire – UKHS (n = 20), UK Large Black – UKLB (n = 20), UK Tamworth – UKTA (n = 20), USA Berkshire – USBK (n = 20), USA Hampshire – USHS (n = 20), and USA Poland China – USPC (n = 6). Sample sizes for all breeds were similar with the exception of the Poland China breed.

Genetic Admixture

The population structure and admixture analyses were performed on the final data set using a Bayesian approach implemented in STRUCTURE software 2.3.4 (Pritchard et al., 2000) without prior information about the population. We had to reduce the number of SNP genotypes in the dataset to 15,000 to enable the complex computations, which otherwise would not have been possible. In order to estimate global ancestry, we used a model with assumed admixture and correlated allele frequencies, as this provides greater power to reveal populations that are closely related (Porras-Hurtado et al., 2013). We performed analyses for the assumed K number of populations from 1 to 34, with 20 independent runs and a burn-in period of 10,000 followed by 100,000 Markov chain Monte Carlo repetitions. The calculations related to STRUCTURE software were performed on the Isabella computer cluster at the University Computing Centre (SRCE) of the University of Zagreb. The choice of the most likely number of clusters (K) was determined according to recommendations in previous work (Pritchard et al., 2000), as well as according to visual representations showing the rate of change in ln Pr(G| K) between successive K-values (Evanno et al., 2005). Clumpak software (Kopelman et al., 2015) was used to estimate the maximum probability from K = 1 until K = 30 and average the individual results among the 20 runs for each K (Jakobsson and Rosenberg, 2007) and over different K-values. The obtained results were visualized using the Pophelper 2.2.7 package for R (Francis, 2017).

ROH and Genomic Inbreeding

The ROH-based genomic inbreeding coefficient (F_{ROH}) was calculated as described (McQuillan et al., 2008; Curik et al., 2014), where F_{ROH} = genome length in ROH/autosomal genome length covered by the SNP chip (here 2,444.5 Mb spread over 18 chromosomes). Based on the SNP density of the Illumina PorcineSNP60 v2 Genotyping BeadChip and the 45,000 SNPs remaining after quality control, ROH were called if 15 or more consecutive homozygous SNPs were present at a density of at

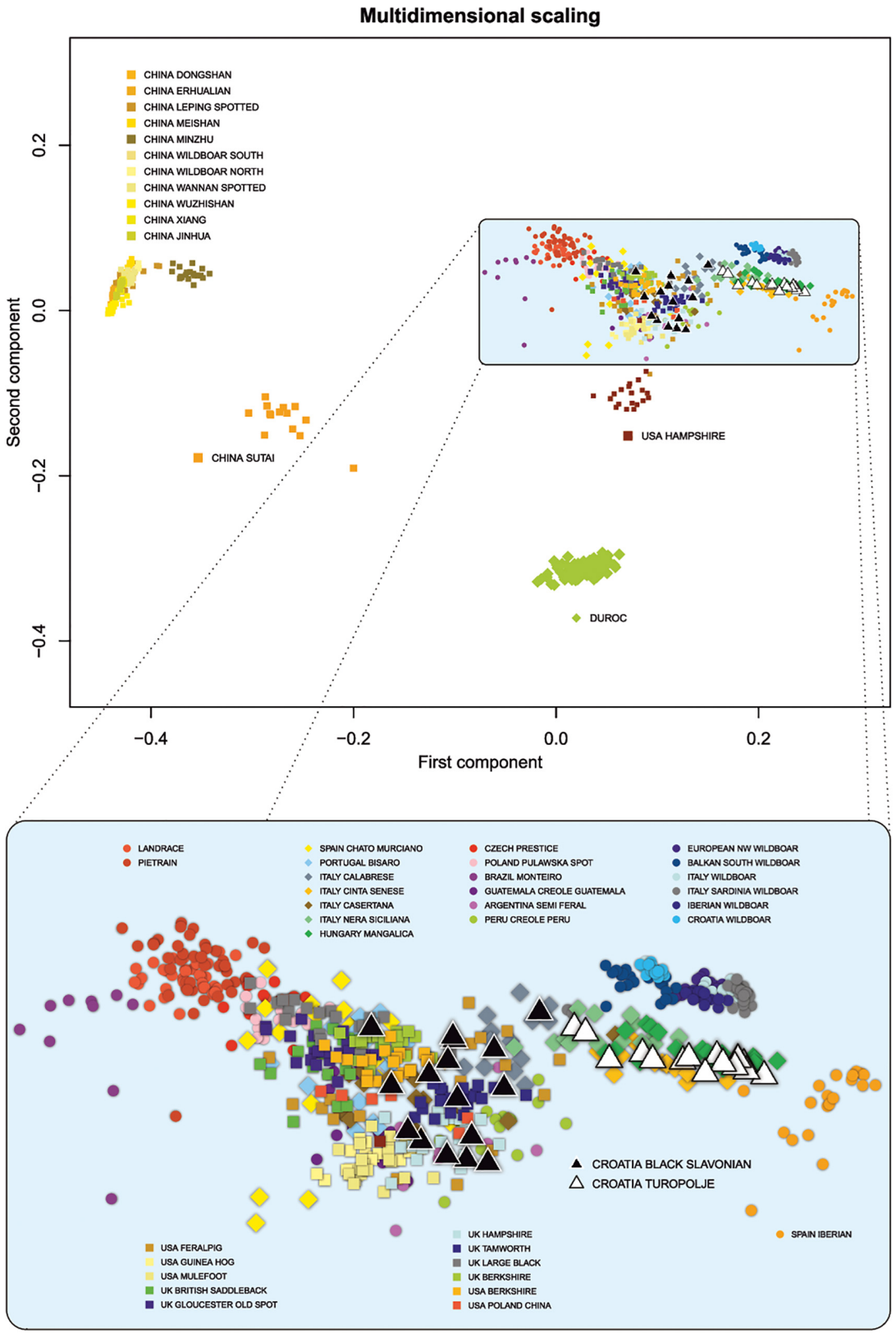


FIGURE 1 | Population structure of the worldwide pig breeds and wild boar populations analyzed by multidimensional scaling (MDS). All breeds/populations are presented in the upper part of the illustration while selection of 32 pig and wild boar populations is presented in the lower part of the illustration.

least one SNP every 0.1 Mb, with gaps of no more than 1 Mb between them. ROH segments were detected using cgaTOH software (Zhang et al., 2013). To identify ROH segments, we allowed one, two and four missing calls per window, respectively, for ROH > 4 Mb, ROH > 8 Mb, and ROH > 16 Mb (Ferenčaković et al., 2013). This approach identifies ROHs according to the length size class. By merging the information related to each class, we were able to calculate genomic inbreeding coefficients ($F_{ROH > 4 Mb}$ and $F_{ROH > 8 Mb}$). Additionally, we calculated $F_{ROH4 to 8 Mb}$ as the difference between $F_{ROH > 4 Mb}$ and $F_{ROH > 8 Mb}$. In this way, we were able to distinguish $F_{ROH>4Mb}$ from "remote" inbreeding ($F_{ROH4to8Mb}$) arising from ancestors approximately –13 generations remote as well as from "recent" inbreeding ($F_{ROH > 8 Mb}$) arising within the last seven generations (Kukučková et al., 2017).

Population Structure and Differentiation of Populations

Global genetic differentiation between the two Croatian local breeds, as well as between the Croatian breeds and other world populations, was assessed in terms of the genome wide fixation index, F_{ST}, for each SNP pair (Weir and Cockerham, 1984). This index was calculated in Plink (Purcell et al., 2007) and GenePop Version 4.7.0 (Rousset, 2008). We also illustrated genetic divergence among breeds/populations by the neighbor-joining tree (NJ) based on Reynold's distances matrix (Reynolds et al., 1983). Reynolds genetic distances were calculated using Arlequin 3.5 (Excoffier and Lischer, 2010) software, which were used to construct a neighbor-joining tree in R package phytools (Revell, 2012).

Identification of Adaptive Signatures of Selection

In order to identify SNP alleles with high F_{ST} values specific to Croatian local breeds, we created two additional datasets, one composed of Black Slavonian and modern commercial breeds (Landrace and Pietrain), and another with Turopolje pigs and the same modern commercial breeds. Based on the two analyses, we selected 30 genome-wide SNPs with the highest F_{ST} values for Black Slavonian and Turopolje pigs (**Supplementary Figures S1, S2** and **Supplementary Tables S2, S3**). The Ensembl Genome Browser[1] was used to identify candidate genes in 0.1 Mb wide genomic regions with high F_{ST} values. In order to explore and confirm the signals of the adaptive positive selection, we performed additional analyses: (a) identification of extremely frequent SNPs in ROHs (eROHi) approach (Curik et al., 2014); (b) extended haplotype homozygosity (EHH) approach (Sabeti et al., 2002) modified as within population Integrated Haplotype Score (iHS) approach (Voight et al., 2006) and (c) across populations Integrated Haplotype Score (Rsb) approach, based on the ratio of site-specific EHH (EHHS) between populations (Tang et al., 2007). Similar approach of combining F_{ST} and extremely frequent ROHs was applied by Purfield et al. (2017). Both, iHS and Rsb statistics were calculated and tested in rehh R package (Gautier and Vitalis, 2012) while required phasing was estimated with Shapeit software (Delaneau et al., 2008). The conservative significance threshold of $P = 0.0001$ (equivalent to 10,000 independent tests), defined with $-\log_{10}$ (P-value) = 4.0, was used in iHS and Rbs tests to account for multiple testing. The eROHi approach has been applied in CROBS, CROTS and commercial pig breeds as our first interest was to identify selection signals that are specific for Croatian local breeds, extreme ROH islands present in CROBS or CROTS but not appearing in commercial breeds. Significant autozygosity islands, SNPs with extreme ROH frequency, were identified as outliers (99%) according to the BOXPLOT distribution as applied in Mészáros et al. (2015). Identified specific regions were then checked for the candidate genes under selection using the free Golden Helix GenomeBrowse® and pig genome assembly Sscrofa 10.2 (EnsEMBL db version 83).

[1]http://www.ensembl.org/Sus_scrofa, Sscrofa 10.2 assembly

RESULTS

Multidimensional Scaling Analysis

In order to analyze the genetic relationship between the Croatian indigenous pig breeds and other worldwide pig breeds or wild boar populations, the MDS approach was used to calculate the shared genetic coancestry among all individuals and breeds/populations (**Figure 1**). Based on the first and second principal components, four main breed clusters were resolved: two Chinese local breed clusters, a Duroc cluster, and a European and North American cluster containing the two indigenous Croatian breeds. The first component clearly separated the Chinese breeds from the European and North American, whereas the second component split one Chinese breed (Sutai), Duroc and Hampshire from the main Chinese and European/North American cluster. A closer look at the European and North American clusters (**Figure 1**, lower part) showed that the commercial breeds Landrace and Pietrain were separated from the breeds in the middle group, which was dominated by the UK and North American local breeds. The wild populations also formed a small independent group, close to the larger group dominated by the Italian local breeds. Croatian indigenous breeds grouped close to the old UK breeds and Italian breeds. As expected, Black Slavonian pigs lay close to its UK and USA breeds of origin: USA Poland China, Berkshire and Large Black. The Turopolje pig breed, in contrast, grouped together with the Italian breeds and Mangalitza, in an intermediate position between the UK local breeds and the wild populations.

Admixture Analysis

The genetic structure of 32 breeds/populations obtained by the STRUCTURE analysis is presented in **Figure 2**, while more detailed explanations about this analysis are provided in **Supplementary Figure S4**. We have presented only results that are relevant for the understanding of the Black Slavonian and Turopolje pig clustering. Thus, the first initial split of $K = 3$ identified a cluster (gray color) belonging to wild populations present also in the Mediterranean and the Pannonian breeds, a cluster (blue) for the European and commercial breeds

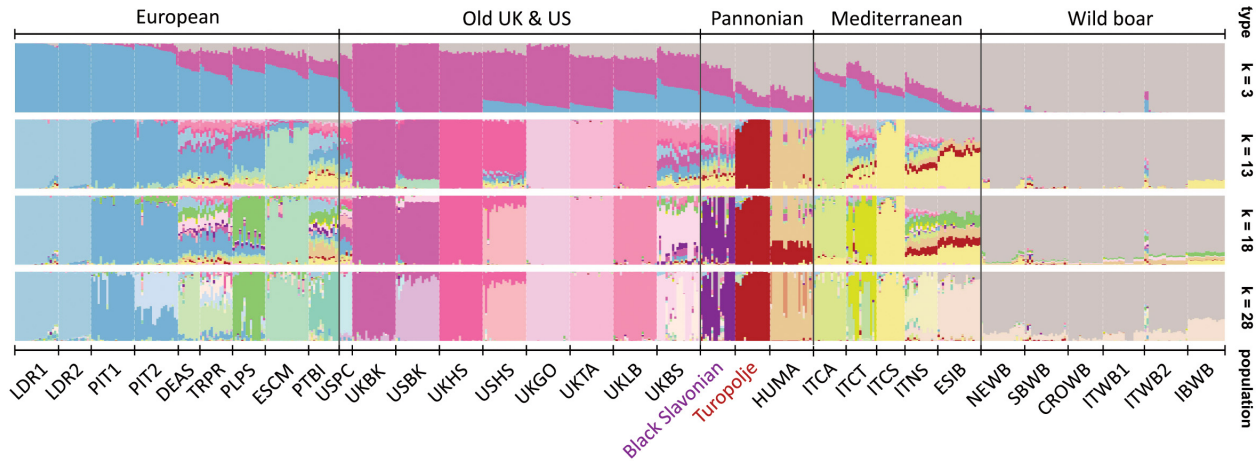


FIGURE 2 | Population structure and admixture for selected 32 pig and wild boar populations illustrated by STRUCTURE bar plots representing models with different clusters (K = 3, K = 13, K = 18, and K = 28). The breeds on the plot are: **Black Slavonian**, Croatia Wild Boar – CROWB, Czech Prestice – TRPR, Germany Angler Sattleschwein – DEAS, Hungarian Mangalitza – HUMA, Iberian Wild Boar – IBWB, Italian Calabrese – ITCA, Italian Casertana –ITCT, Italian Cinta Senese –ITCS, Italian Nera Siciliana –ITNS, Italian Sardinian Wild Boar – ITWB2, Italian Wild Boar – ITWB1, Landrace population 1 – LDR1, Landrace population 2 – LDR2, NW European Wild Boar –NEWB, Pietrain population 1 – PIT1, Pietrain population 2 – PIT2, Poland Pulawska Spot – PLPS, Portugal Bisaro –PTBI, South Balkan Wild Boar – SBWB, Spain Chato Murciano –ESCM, Spanish Iberian –ESIB, **Turopolje**, UK Berkshire –UKBK, UK British Saddleback –UKBS, UK Gloucester Old Spot –UKGO, UK Hampshire –UKHS, UK Large Black –UKLB, UK Tamworth – UKTA, USA Berkshire – USBK, USA Hampshire – USHS, and USA Poland China – USPC.

influencing the Mediterranean more than the Pannonian breeds, and a cluster (pink) for the old UK and US breeds influencing the European and Pannonian breeds more than the Mediterranean ones. At K = 13, the Turopolje pig constituted a unique cluster whereas the Black Slavonian breed was identified as a single cluster only from K = 18, while it showed high genetic admixture with modern breeds. Despite their geographical proximity, we did not observe any admixture traces between Black Slavonian and Turopolje pigs. In the most likely model of K = 28, most of the 24 breeds appeared as individual clusters, except for German Angler Sattleschwein (DEAS) and Czech Prestice (TRPR), which overlapped. All six wild boar populations appeared as a separate group across all K values. However, at K = 28, a small amount of the Spanish Iberian (ESIB) component was present in all wild boar populations, particularly in the Iberian Wild Boar (IBWB). At the low level of differentiation (K = 3), the largest amount of the wild boar cluster was present in the Mediterranean and Pannonian breeds. One Pietrain population (PIT2), UK Berkshire (UKBS), Hungarian Mangalitza (HUMA) and ITCT (Italian Casertana) showed slight sub-structuring, while the USA Berkshire (USBK) and USA Hampshire (USHS) breeds appeared as more separated in comparison to the other breeds from UK and USA (**Figure 2**).

ROH-Based Analysis of Genomic Inbreeding

The distribution of the ROH inbreeding coefficients ($F_{ROH\ >\ 4\ Mb}$) for all the analyzed breeds and wild populations is presented in **Figure 3**. The same figure also shows the contribution (%) of the "close" inbreeding ($F_{ROH>\ 8\ Mb}$) caused by ancestors within seven generations relative to the total inbreeding level ($F_{ROH\ >\ 4\ Mb}$). Among the considered pig breeds/populations, 65–89% of inbreeding came from 'close' inbreeding. "Close" inbreeding was significantly lower ($P < 0.001$) in wild boar individuals (mean $F_{ROH\ >\ 8\ Mb}$ = 0.112; 95% confidence interval, CI = 0.094–0.128) than in domestic breeds (mean $F_{ROH\ >\ 8\ Mb}$ = 0.172; 95% CI = 0.162–0.182). The difference remained significant ($P < 0.001$) even when two populations with high outlying inbreeding were excluded from the analysis (mean $F_{ROH>8Mb}$ = 0.156; 95% CI = 0.147–0.165). This was unexpected, since domestic pigs are bred based on pedigree information in order to avoid mating of relatives within six to seven generations. Extremely high inbreeding values ($F_{ROH\ >\ 4\ Mb}$ = 0.400 and $F_{ROH>8Mb}$ = 0.332; **Supplementary Figure S3**) were observed in Turopolje pig, and such extreme inbreeding values were observed only in Hungarian Mangalitza (HUMA) ($F_{ROH\ >\ 4\ Mb}$ = 0.415, and $F_{ROH\ >\ 8\ Mb}$ = 0.371). An increased frequency of very long ROH (>30 Mb), showing increased close inbreeding, was also observed in Romanian and Hungarian Red Mangalitza pigs (Bâlteanu et al., 2019). In contrast, much lower inbreeding was observed in Black Slavonian pigs ($F_{ROH\ >\ 4\ Mb}$ = 0.098 and $F_{ROH\ >\ 8\ Mb}$ = 0.074).

Analysis of Population Structure and Differentiation of Populations

The population differentiation was analyzed by pairwise F_{ST} values estimated across all populations from the final dataset of 32 breeds/populations (**Supplementary Table S4**). The mean F_{ST} estimate was 0.25, while all pairwise F_{ST} values from a selection of breeds are shown in **Table 1**. The F_{ST} values ranged from 0.07 (between the two Pietrain populations) to 0.40 (between Turopolje and Gloucester Old Spot pig breed). Genetic differentiation tends to be smaller between local pig breeds with a closer genetic history. The Black Slavonian breed showed a low

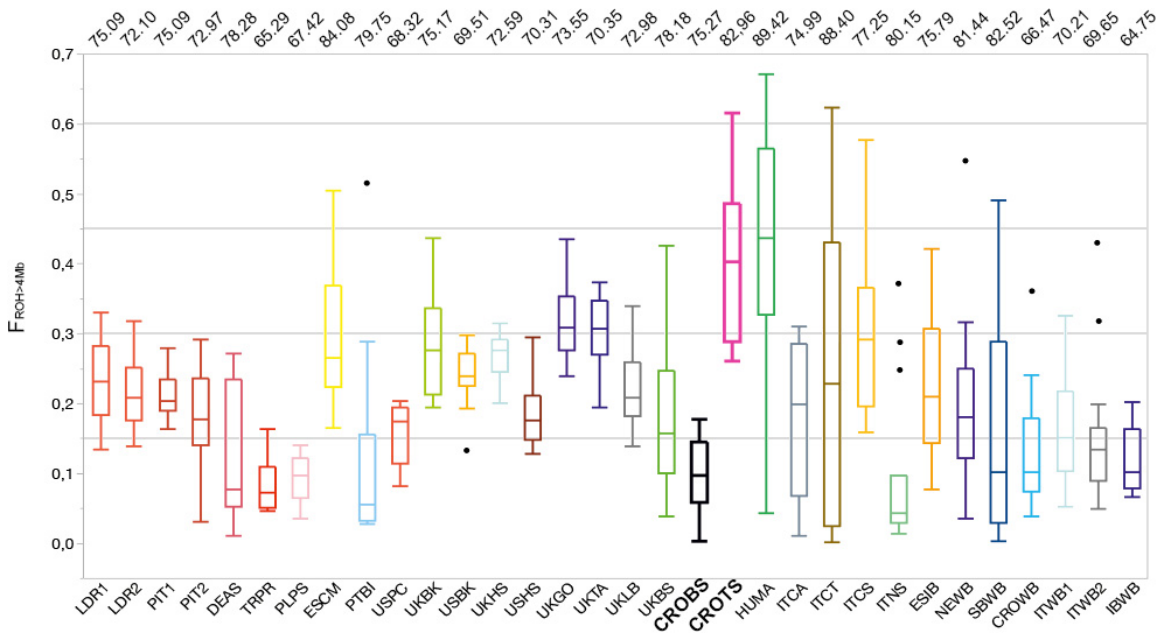


FIGURE 3 | The distribution of the ROH inbreeding coefficients ($F_{ROH\ >\ 4\ Mb}$) for selected 32 pig and wild boar populations. Numbers on the top of the illustration present the contribution (%) of the "close" inbreeding ($F_{ROH\ >\ 8\ Mb}$) caused by ancestors within seven generations relative to the total inbreeding level ($F_{ROH\ >\ 4\ Mb}$). The breeds on the plot are: **Black Slavonian – CROBS**, Croatian Wild Boar – CROWB, Czech Prestice – TRPR, German Angler Sattleschwein – DEAS, Hungarian Mangalitza – HUMA, Iberian Wild Boar – IBWB, Italian Calabrese – ITCA, Italian Casertana – ITCT, Italian Cinta Senese – ITCS, Italian Nera Siciliana – ITNS, Italian Sardinian Wild Boar – ITWB2, Italian Wild Boar – ITWB1, Landrace population 1 – LDR1, Landrace population 2 – LDR2, NW European Wild Boar – NEWB, Pietrain population 1 – PIT1, Pietrain population 2 – PIT2, Polish Pulawska Spot – PLPS, Portuguese Bisaro – PTBI, South Balkan Wild Boar – SBWB, Spanish Chato Murciano – ESCM, Spanish Iberian – ESIB, **Turopolje – CROTS**, UK Berkshire – UKBK, UK British Saddleback – UKBS, UK Gloucester Old Spot – UKGO, UK Hampshire – UKHS, UK Large Black – UKLB, UK Tamworth – UKTA, USA Berkshire – USBK, USA Hampshire – USHS, and USA Poland China – USPC.

mean F_{ST} value (0.21), consistent with its central position in the international dataset, while Turopolje pig had higher mean F_{ST} value (0.32), consistent with its peripheral position. The highest mean F_{ST} estimate among all breeds in this study was 0.33 for UK Tamworth; the lowest estimate was 0.18 for Czech Prestice. The observed values of genetic differentiation are comparable to the results of other studies.

Genetic differentiation among 32 breeds/populations was further illustrated by unrooted neighbor-joining tree based on Reynolds genetic distances (**Figure 4**). The presented tree clearly shows differentiation of wild boar populations from commercial breeds while a number of indigenous breeds are presented between this separation of two extreme groups (wild boar populations versus commercial breeds) in a succeeding manner, starting with Iberian pig (ESIB) as the closest breed to the wild populations and ending with Black Slavonian breed as the closest breed to the commercial breeds. In this separation route Turopolje pig is positioned in the middle.

Identification of Adaptive Signatures of Selection

In order to identify loci specific to Croatian indigenous breeds and therefore more useful for conservation efforts, we separately calculated the locus-wise F_{ST} values between Croatian local breeds and modern pig breeds (Landrace and Pietrain). We selected 30 genome-wide SNPs with the highest F_{ST} values for Black Slavonian and Turopolje pigs (**Supplementary Figures S1, S2**). Genes located within the genomic regions of SNPs with extremely high F_{ST} were identified as candidate genes that could help in future conservation programs. Most likely polymorphisms in these genes are the consequence of breed adaptation to environmental and human demands. For Black Slavonian pig, we identified important genes associated with steroid receptor activity, such as CYP-40 on porcine chromosome SSC8 (Ratajczak et al., 2015); meat-to-fat ratio in pigs, DEAF1 on SSC2 (Falker-Gieske et al., 2019); growth traits in cattle, KSR2 on SSC14 (Puig-Oliveras et al., 2014); animal organ and system development in pigs, SEZ6L on SSC14 (Kwon et al., 2019); hematological parameters in pigs, RHOBTB1 on SSC 14 (Bovo et al., 2019); female reproduction in mice, CDK1 on SSC 14 (Adhikari et al., 2016); salivary secretion in pigs, KCNMA1 on SSC14 (Li et al., 2013); milk fat percentage in buffaloes, KCTD8 on SSC8 (de Camargo et al., 2015); back fat thickness in pigs, RIMS4 on SSC17 (Lee and Shin, 2018); carcass length in pigs, SPTLC2 on SSC7 (Falker-Gieske et al., 2019); muscle fiber types in pigs, MYO18B on SSC14 (Ropka-Molik et al., 2018); and fatty acid profiles in cattle, RAPGEF2 on SSC8 (Cesar et al., 2014).

For Turopolje pig, we identified candidate genes associated with fatty acid metabolism in pigs, such as PEX11A on SSC7 (Huang et al., 2017); carcass traits in cattle, WDR93 on SSC7

TABLE 1 | Genetic differentiation among pig breeds/populations based on F_{ST} estimates.

Breed/population	DEAS	ITCS	ITCT	ESIB	UKLB	LDR1	HUMA	NEWB	SBWB	CROWB	TRPR	USPC	UKBK	CROBS	F_{ST}
Italy Cinta Senese – ITCS	0.25														0.29
Italy Casertana – ITCT	0.18	0.25													0.23
Spain Iberian – ESIB	0.21	0.22	0.19												0.23
UK Large Black – UKLB	0.21	0.30	0.23	0.25											0.26
Landrace – LDR1	0.19	0.30	0.23	0.28	0.27										0.27
Hungary Mangalica – HUMA	0.28	0.31	0.27	0.22	0.30	0.33									0.29
NW European Wild Boar – NEWB	0.26	0.28	0.25	0.19	0.29	0.31	0.28								0.26
South Balkan Wild Boar – SBWB	0.24	0.27	0.22	0.18	0.27	0.30	0.26	0.13							0.24
Croatia Wild Boar – CROWB	0.27	0.30	0.25	0.21	0.29	0.32	0.29	0.15	0.10						0.26
Czech Prestice – TRPR	0.08	0.21	0.15	0.17	0.17	0.18	0.23	0.22	0.20	0.22					0.18
USA Poland China – USPC	0.19	0.30	0.22	0.24	0.25	0.26	0.32	0.29	0.27	0.30	0.15				0.25
UK Berkshire – UKBK	0.25	0.35	0.27	0.29	0.29	0.30	0.34	0.33	0.31	0.34	0.21	0.28			0.29
Black Slavonian pig – CROBS	0.17	0.24	0.18	0.18	0.20	0.23	0.24	0.22	0.20	0.23	**0.13**	0.21	0.24		0.21
Turopolje pig – CROTS	0.30	0.35	0.29	0.26	0.33	0.34	0.34	0.32	0.30	0.33	**0.25**	0.35	0.37	0.28	0.32

Pairwise F_{ST} estimates (Weir and Cockerham, 1984) presenting genetic differentiation among 14 selected breeds/populations and average F_{ST} estimates for each breed/population (right column).

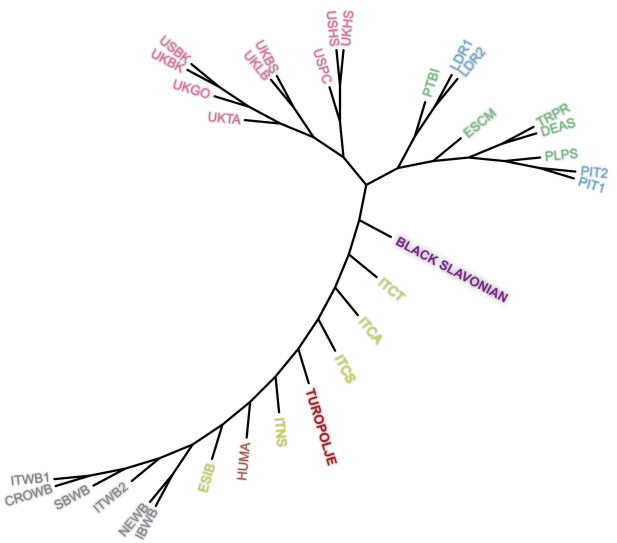


FIGURE 4 | Neighbor-joining tree based on the Reynold's genetic distances for selected 32 pig and wild boar populations. The breeds on the plot are: **Black Slavonian**, Croatian Wild Boar – CROWB, Czech Prestice – TRPR, German Angler Sattleschwein – DEAS, Hungarian Mangalitza – HUMA, Iberian Wild Boar – IBWB, Italian Calabrese – ITCA, Italian Casertana – ITCT, Italian Cinta Senese – ITCS, Italian Nera Siciliana – ITNS, Italian Sardinian Wild Boar – ITWB2, Italian Wild Boar – ITWB1, Landrace population 1 – LDR1, Landrace population 2 – LDR2, NW European Wild Boar – NEWB, Pietrain population 1 – PIT1, Pietrain population 2 – PIT2, Polish Pulawska Spot – PLPS, Portuguese Bisaro – PTBI, South Balkan Wild Boar – SBWB, Spanish Chato Murciano – ESCM, Spanish Iberian – ESIB, **Turopolje**, UK Berkshire – UKBK, UK British Saddleback – UKBS, UK Gloucester Old Spot – UKGO, UK Hampshire – UKHS, UK Large Black – UKLB, UK Tamworth – UKTA, USA Berkshire – USBK, USA Hampshire – USHS, and USA Poland China – USPC.

(Silva et al., 2017); number of ribs in pigs, MESP1 on SSC7 (Zhu et al., 2015); meat-to-fat ratio in pigs, DEAF1 on SSC2 (Falker-Gieske et al., 2019); pregnancy rate in pigs, PPID on SSC8 (Gu et al., 2014); steroid receptor activity, CYP-40 on SSC8 (Ratajczak et al., 2015); brain development in horses, DLGAP1 on SSC6 (Schubert et al., 2014); salivary secretion in pigs, KCNMA1 on SSC14 (Li et al., 2013); reproduction in pigs, CWH43 on SSC8 (He et al., 2017); bone weight in cattle, FAM184B on SSC8 (Xia et al., 2017) and cardiovascular disease (Pérez-Montarelo et al., 2014); spermiogenesis in mouse, AMPH on SSC9; boar taint, NWD2 on SSC8 (Drag et al., 2018); melanocyte function in dogs, ARHGAP12 on SSC10 (Kluth and Distl, 2013); female pregnancy in pigs, RAPGEF2 on SSC8 (Pérez-Enciso et al., 2009); growth traits in cattle, KSR2 on SSC14 (Puig-Oliveras et al., 2014); vascular smooth muscle contraction in sheep, SPSB4 on SSC13 (Yang et al., 2016); and back-fat fatty acid composition, APBB1IP on SSC10 (Zappaterra et al., 2018).

In addition, we identified several SNPs in the two Croatian breeds that were located in non-coding intergenic regions and that were present in various pig breeds as well as to other domestic animal species, including cattle, sheep and horse. In Black Slavonian pig, the following 12 SNPs were identified: ASGA0012664, ALGA0082391, SIRI0000509, ALGA0115258, ALGA0008072, ASGA0038761, ASGA0080338,

ALGA0098790, ASGA0039781, ASGA0039779, M1GA0015147, and MARC0003342. In Turopolje pig, the following nine SNPs were identified: MARC0067231, ALGA0115258, M1GA0015147, ASGA0038761, ALGA0048121, MARC0085941, ASGA0042725, ASGA0038765, and SIRI0000509.

To provide additional support to the identification of genome regions with adaptive selection signatures, we also performed several tests that are used in the identification of selection signatures such as eROHi, iHS and Rsb analysis. The overall results of the selection signature analyses are presented in **Supplementary Table S5**. Among all approaches performed, F_{ST} and Rsb analyses are the most similar by the concept as they are both looking for genome segments that are selected in indigenous breeds in contrast to commercial populations, while eROHi and iHS analyses are based on the identification of adaptive selection signatures from genomic information of the single population.

We have not identified any significant SNP overlapping between F_{ST} and Rsb analyses neither in Black Slavonian nor in Turopolje pig population. However, when we were looking for the overlapping results between F_{ST} and eROHi analysis, three significant SNPs (MARC0058238, MARC0003342, and ALGA0077279, all on SSC14) pointing to the adaptive selection signals, were observed in Black Slavonian breed while only one such significant SNP (ALGA0036219 on SSC6) was observed in Turopolje breed. The first SNP for Black Slavonian was previously described (MARC0058238, located in the MYO18B genomic region on SSC14, which is found to be associated with muscle fiber types in pigs- Ropka-Molik et al., 2018). Second (MARC0003342) and third (ALGA0077279) SNP identified in Black Slavonian pig are located in non-coding intergenic region present in various domestic animal species, together with the SNP (ALGA0036219 on SSC6) found in Turopolje pig. We identified one additional SNP (ASGA0060892 on SSC14) with significant selection signal obtained in both eROHi and iHS analyses in Turopolje pig. This variant, located in PEBP4 gene region, is associated with hematological traits in pigs (Bovo et al., 2019) and has been shown to differentiate Chinese local breeds from Large White pigs (Li et al., 2014).

DISCUSSION

Over the last hundred years, strong demand for animal protein and economic efficiency, combined with globalization and market competition, have intensified pig breeding and selection, leading to the domination of several commercial breeds such as Large White, Duroc, Landrace, Hampshire, and Pietrain. In the last few decades, many valuable local breeds have gone extinct or are on the brink of extinction. Conserving these species is important for maintaining genetic diversity to promote long-term selection progress (Bruford et al., 2015).

Black Slavonian and Turopolje pigs are Croatian local indigenous breeds that are well adapted to harsh environments and should be preserved from extinction as they can contribute to the overall adaptive genetic potential. In this study, based on high-throughput genomic information, Black Slavonian and Turopolje pig breeds were genetically compared with many internationally relevant breeds, as well as with several wild boar populations. MDS multivariate analysis and unsupervised clustering showed that both breeds have complex but close genetic relatedness with other European pig breeds, and can be considered part of the living European livestock (pig) heritage (**Figures 1** and **2**). The Black Slavonian pig appears to be more influenced by the classical West European breeds, while the Turopolje pig clusters with the Mediterranean pig breeds in vicinity to the cluster representing wild boars (**Figures 1**, **2**). Still, the algorithm implemented in our STRUCTURE analysis was able to make a distinction among Turopolje pig (at $K = 13$), Black Slavonian pig (at $K = 18$), and other European pig breeds. Turopolje pig showed a low level of admixture with commercial pigs, while Black Slavonian pig showed greater and more variable admixture (**Figure 2**). The admixture contributions in the Black Slavonian pig originated from several equally contributing clusters belonging to different commercial breeds. We speculate that these are signals of admixture with some of the modern pure breed or hybrid pigs commercially reared in Slavonia, pointing to the need for further maintenance of systematic breeding programs for breed consolidation and recovery. High inbreeding values, particularly the recent ones, were obtained for the Turopolje pig. With the exception of the Hungarian Mangalitza breed, such high inbreeding values have not been reported for the other breeds analyzed here (Saura et al., 2013; Schäler et al., 2020). The observed values exceed considerably even the expected inbreeding that would result from full sib or parent-offspring mating, and they seriously threaten the survival of the breed. A much better situation, with a relatively low inbreeding level, was observed in Black Slavonian pig, even if the breed went through a severe bottleneck in the 1990's. The presence of admixture signals could certainly have an impact on the observed inbreeding level, but only for the admixed individuals. Thus, we think that the observed inbreeding level is the consequence of the recent breeding program and pedigree-controlled mating strategy performed in the last decade, according to which only sows and boars with known ancestry and acceptable coefficient of relationship were allowed to mate.

A recent study analyzed genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds, including Black Slavonian and Turopolje pig (Muñoz et al., 2019). Their aims were slightly different from those of the present study, and their analyses were oriented toward European local breeds more generally. In contrast, we were interested in conservation genomics and estimation of admixture and genomic inbreeding in the two indigenous Croatian breeds. Thus, their analysis relied on GeneSeek® GGP Porcine HD Genomic Profiler v1 markers, while ours relied on PorcineSNP60 v2 markers. We were aware that the sample size in our study was small for the reliable estimation of gametic (Ne_{GD}) or/and linkage disequilibrium (Ne_{LD}) effective population size. Their analysis estimated very small effective population size based on linkage disequilibrium (N_{eLD}) in Black Slavonian pigs ($N_{eLD} = 33$) and Turopolje pigs ($N_{eLD} = 10$) for the current generation, although the estimates of the contemporary N_{eLD} population size are quite sensitive (Corbin et al., 2012). Future work, on a larger sample size should estimate these parameters because they are important for conservation assessment of these Croatian indigenous breeds.

Nevertheless, the sample sizes of breeds or populations in the present study are comparable to those in similar studies (Burgos-Paz et al., 2013; Goedbloed et al., 2013; Decker et al., 2014) and were appropriate for analyses on MDS, population structure and admixture with the STRUCTURE algorithm, estimation of genomic inbreeding and identification of breed-specific genome regions. A larger sample size would narrow the CI of the estimated inbreeding level but not alter our conclusions, since the estimated marginal values of the confidence intervals in Turopolje pig were extremely high (the 95% CI was 0.342–0.459 for $F_{ROH > 4\ Mb}$, and the 95% CI was 0.280–0.383 for $F_{ROH > 8\ Mb}$), whereas those obtained in Black Slavonian pig were relatively low (the 95% CI was 0.071–0.123 for $F_{ROH > 4\ Mb}$, and the 95% CI was 0.050–0.097 for $F_{ROH > 8\ Mb}$) compared to the estimated inbreeding level in other breeds in the present study as well as in other studies (Saura et al., 2013; Schäler et al., 2020).

In addition, it is important to highlight that Turopolje pig population consists of 124 sows and 17 boars (Croatian Agricultural Agency, 2017). Also, the ascertainment bias could have influenced the analyses performed, since both Black Slavonian and Turopolje pigs are local breeds that were not included in the development of the Illumina PorcineSNP60 v2 Genotyping BeadChip. However, such influence is likely to be minimal while our findings should be verified in studies based on whole-genome sequencing.

We obtained additional insights into the genetic background of Croatian local pigs through the identification of genomic regions that show a high level of differentiation (extreme F_{ST}) between the Croatian indigenous pigs and commercial modern animals. Genes identified within those regions are likely to have important adaptive functions and therefore are suitable for traceability studies to protect and promote products derived from Black Slavonian and Turopolje pig. In addition to their expected functions, these candidate genes have been associated with production or carcass traits in Black Slavonian pig (DEAF1, KSR2, RIMS4, and SPTLC2) and Turopolje pig (WDR93, MESP1, DEAF1, and KSR2), as well as with reproduction and system development in Black Slavonian pig (SEZ6L, RHOBTB1, CDK1, and KCNMA1) and Turopolje pig (PPID, KCNMA1, CWH43, RAPGEF2, SPSB4, and APBB1IP). In identifying adaptive selection signals with F_{ST} analysis we were extremely conservative as only 30 SNPs with highest F_{ST} values were considered significant. We wanted to minimize the number of false positive selection signatures. Thus, we have performed additional analyses toward identification of selection signals (eROHi, iHS, and Rsb). The presence of selection signatures obtained by F_{ST} analysis was confirmed for three genome regions in Black Savonian breed and one genome region in Turopolje breed. Additional selection signature has been identified in PEBP4 gene region (placed on SSC14) in Turopolje breed as significant signals were obtained by eROHi and iHS analyses. Assuming that the necessary data become available, future work may wish to take a more comprehensive approach, at least for the relatively large Black Slavonian population, by estimating breeding values for traits of conservation interest and combining those estimates with F_{ST} values to detect conservation-relevant SNPs (Zhang et al., 2014).

CONCLUSION

In conclusion, our results show that Black Slavonian and Turopolje pigs are distinct breeds genetically related to other European pig breeds. Uncontrolled breeding is likely to reduce the genomic diversity of European pig breeding capacity and threaten the cultural heritage of these breeds. Although conservation planning has already been implemented for the Black Slavonian pig, and our results suggest that such planning has benefited the breed, future actions toward admixture consolidation and management are required. The conservation status of the Turopolje pig is alarming and an urgent conservation plan is needed. The two local breeds in this study currently make only a marginal contribution to commercial pig production, yet we need to protect the genetic variability of these local breeds to guarantee necessary genetic diversity for the future. The identification of breed specific genome regions with extreme F_{ST} values will enable protection and promotion of commercial products derived from Black Slavonian and Turopolje pigs.

DATA AVAILABILITY STATEMENT

This manuscript contains previously unpublished data available at Dryad (https://doi.org/10.5061/dryad.wpzgmsbhq).

ETHICS STATEMENT

All procedures used for this study involving animals were in accordance with the guidelines for animal welfare defined by the Croatian Ministry of Agriculture.

AUTHOR CONTRIBUTIONS

VC-C and IC conceived the research. MČ, VO, and LI conducted the field work and laboratory analyses. BL, MF, and DŠ analyzed the data. BL, VC-C, and IC wrote the manuscript. All authors drafted and approved the current version of the manuscript.

FUNDING

This study was supported by the Croatian Science Foundation (project ANAGRAMS-IP-2018-01-8708). LI received funding from the Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Action (Grant Agreement no. 656697).

ACKNOWLEDGMENTS

We wish to acknowledge Danijel Vrčić, Emir Imamagić and Tomislav Stilinović at the University Computing Centre (SRCE) of the University of Zagreb for providing computational facilities (Isabella cluster) and support. We also thank Zlatko Šatović and

Ante Turudić for sharing their experience with the STRUCTURE software in the Isabella cluster, and Ras Lužaić for the graphical enhancements of the figures. Publication was supported by the OpenAccess Publication Fund of the University of Zagreb Faculty of Agriculture.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00261/full#supplementary-material

REFERENCES

Adhikari, D., Busayavalasa, K., Zhang, J., Hu, M., Risal, S., Bayazit, M. B., et al. (2016). Inhibitory phosphorylation of Cdk1 mediates prolonged prophase I arrest in female germ cells and is essential for female reproductive lifespan. *Cell Res.* 26, 1212–1225. doi: 10.1038/cr.2016.119

Ai, H., Huang, L., and Ren, J. (2013). Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western Pigs revealed by genome-wide SNP markers. *PLoS One* 8:e56001. doi: 10.1371/journal.pone.0056001

Balding, D. J., Bishop, M., and Cannings, C. (2007). *Handbook of Statistical Genetics, Vol. 1*. West Sussex, England: John Wiley & Sons, Ltd.

Bâlteanu, V. A., Cardoso, T. F., Amills, M., Egerszegi, I., Anton, I., Beja-Pereira, A., et al. (2019). The footprint of recent and strong demographic decline in the genomes of Mangalitza pigs. *Animal* 5, 1–7. doi: 10.1017/S1751731119000582

Bovo, S., Mazzoni, G., Bertolini, F., Schiavo, G., Galimberti, G., Gallo, M., et al. (2019). Genome-wide association studies for 30 haematological and blood clinical-biochemical traits in large white pigs reveal genomic regions affecting intermediate phenotypes. *Sci. Rep.* 9:7003. doi: 10.1038/s41598-019-43297-1

Bruford, M. W., Ginja, C., Hoffmann, I., Joost, S., Orozco-ter Wengel, P., Alberto, F. J., et al. (2015). Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Front. Genet.* 6:314. doi: 10.3389/fgene.2015.00314

Burgos-Paz, W., Souza, C. A., Megens, H. J., Ramayo-Caldas, Y., Melo, M., Lemús-Flores, C., et al. (2013). Porcine colonization of the Americas: a 60k SNP story. *Heredity* 110, 321–330. doi: 10.1038/hdy.2014.81

Cesar, A. S., Regitano, L. C., Mourão, G. B., Tullio, R. R., Lanna, D. P., Nassu, R. T., et al. (2014). Genome-wide association study for intramuscular fat deposition and composition in Nellore cattle. *BMC Genet.* 15:39. doi: 10.1186/1471-2156-15-39

Corbin, L. J., Liu, A. Y. H., Bishop, S. C., and Wooliams, J. A. (2012). Estimation of historical effective population size using linkage disequilibrium with marker data. *J. Anim. Breed. Genet.* 129, 257–270. doi: 10.1111/j.1439-0388.2012.01003.x

Croatian Agricultural Agency, (2017). *Annual Report for Pig Breeding 2018*. Križevci: Croatian Agricultural Agency.

Curik, I., Ferenčaković, M., and Sölkner, J. (2014). Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livest. Sci.* 166, 26–34. doi: 10.1016/j.livsci.2014.05.034

de Camargo, G. M. F., Aspilcueta-Borquis, R. R., Fortes, M. R. S., Porto-Neto, R., Cardoso, D. F., Santos, D. J. A., et al. (2015). Prospecting major genes in dairy buffaloes. *BMC Genomics* 16:872. doi: 10.1186/s12864-015-1986-2

Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J., Molina Alcalá, A., Sonstegard, T. S., et al. (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.* 10:e1004254. doi: 10.1371/journal.pgen.1004254

Delaneau, O., Coulonges, C., and Zagury, J. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 9:540. doi: 10.1186/1471-2105-9-540

Drag, M., Hansen, M. B., and Kadarmideen, H. N. (2018). Systems genomics study reveals expression quantitative trait loci, regulator genes and pathways associated with boar taint in pigs. *PLoS One* 13:e0192673. doi: 10.1371/journal.pone.0192673

Druml, T., Salajpal, K., Dikic, M., Urosevic, M., Grilz-Seger, G., and Baumung, R. (2012). Genetic diversity, population structure and subdivision of local Balkan pig breeds in Austria, Croatia, Serbia and Bosnia-Herzegovina and its practical value in conservation programs. *Genet. Sel. Evol.* 44:5. doi: 10.1186/1297-9686-44-5

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 ver 3.5: 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

Falker-Gieske, C., Blaj, I., Preuß, S., Bennewitz, J., Thaller, G., and Tetens, J. (2019). GWAS for meat and carcass traits using imputed sequence level genotypes in pooled F2-designs in Pigs. *G3: Genes Genom. Genet.* 9, 2823–2834. doi: 10.1534/g3.119.400452

FAO, (2007). *The State of the World's Animal Genetic Resources for Food and Agriculture*, eds B. Rischkowsky, and D. Pilling, (Rome: FAO).

Ferenčaković, M., Sölkner, J., and Curik, I. (2013). Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genet. Sel. Evol.* 45:42. doi: 10.1186/1297-9686-45-42

Francis, R. M. (2017). POPHELPER: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* 17, 27–32. doi: 10.1111/1755-0998.12509

Frantz, L., Haile, J., Lin, A. T., Scheu, A., Geörg, C., Benecke, N., et al. (2019). Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proc. Natl. Acad. Sci. U.S.A.* 116, 17231–17238. doi: 10.1073/pnas.1901169116

Gautier, M., and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28, 1176–1177. doi: 10.1093/bioinformatics/bts115

Goedbloed, D. J., Megens, H. J., Van Hooft, P., Herrero-Medrano, J. M., Lutz, W., Alexandri, P., et al. (2013). Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into northwest European wild boar populations. *Mol. Ecol.* 22, 856–866. doi: 10.1111/j.1365-294X.2012.05670.x

Gu, T., Zhu, M. J., Schroyen, M., Qu, L., Nettleton, D., Kuhar, D., et al. (2014). Endometrial gene expression profiling in pregnant Meishan and Yorkshire pigs on day 12 of gestation. *BMC Genomics* 15:156. doi: 10.1186/1471-2164-15-156

He, L. C., Li, P. H., Ma, X., Sui, S. P., Gao, S., Kim, S. W., et al. (2017). Identification of new single nucleotide polymorphisms affecting total number born and candidate genes related to ovulation rate in Chinese Erhualian pigs. *Anim. Genet.* 48, 48–54. doi: 10.1111/age.12492

Hrasnica, F., Ilančić, D., Pavlović, S., Rako, A., and Šmalcelj, I. (1958). *Specijalno Stočarstvo*. Zagreb: Poljoprivredni nakladni zavod.

Huang, W., Zhang, X., Li, A., Xie, L., and Miao, X. (2017). Differential regulation of mRNAs and lncRNAs related to lipid metabolism in two pig breeds. *Oncotarget* 8, 87539–87553. doi: 10.18632/oncotarget.20978

Iacolina, L., Pertoldi, C., Amills, M., Kusza, S., Megens, H. J., Bâlteanu, V. A., et al. (2018). Hotspots of recent hybridization between pigs and wild boars in Europe. *Sci. Rep.* 8:17372. doi: 10.1038/s41598-018-35865-8

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Keller, M. C., Visscher, P. M., and Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237–249. doi: 10.1534/genetics.111.130922

Kluth, S., and Distl, O. (2013). Congenital sensorineural deafness in Dalmatian dogs associated with quantitative trait loci. *PLoS One* 8:e80642. doi: 10.1371/journal.pone.0080642

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* 15, 1179–1191. doi: 10.1111/1755-0998.12387

Kukučková, V., Moravčíková, N., Ferenčaković, M., Simčič, M., Mészáros, G., Sölkner, J., et al. (2017). Genomic characterization of Pinzgau cattle: genetic conservation and breeding perspectives. *Conserv. Genet.* 18, 893–910. doi: 10.1007/s10592-017-0935-9

Kwon, D. J., Lee, Y. S., Shin, D., Won, K. H., and Song, K. D. (2019). Genome analysis of Yucatan miniature pigs to assess their potential as biomedical model animals. *Asian-Australas. J. Anim. Sci.* 32, 290–296. doi: 10.5713/ajas.18.0170

Lee, Y. S., and Shin, D. (2018). Genome-Wide association studies associated with backfat thickness in Landrace and Yorkshire Pigs. *Genomics Inform.* 16, 59–64. doi: 10.5808/GI.2018.16.3.59

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., et al. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* 45, 1431–1438. doi: 10.1038/ng.2811

Li, X., Yang, S., Tang, Z., Li, K., Rothschild, M. F., Liu, B., et al. (2014). Genome-wide scans to detect positive selection in large white and Tongcheng pigs. *Anim. Genet.* 45, 329–339. doi: 10.1111/age.12128

Lukić, B., Smetko, A., Mahnet, Ž, Klišanić, V., Špehar, M., Raguž, N., et al. (2015). Population genetic structure of autochthonous black Slavonian Pig. *Agriculture* 21, 28–32. doi: 10.18047/poljo.21.1.sup.5

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., et al. (2008). Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 83, 359–372. doi: 10.1016/j.ajhg.2008.08.007

Mészáros, G., Boison, S. A., Pérez O'Brian, A. M., Ferenčaković, M., Curik, I., da Silva, M. V. B., et al. (2015). Genomic analysis for managing small and endangered populations: a case study in Tyrol Grey cattle. *Front. Genet.* 6:173. doi: 10.3389/fgene.2015.00173

Muñoz, M., Bozzi, R., García, F., Núñez, Y., Geraci, C., Crovetti, A., et al. (2018). Diversity across major and candidate genes in European local pig breeds. *PLoS One* 13:e0207475. doi: 10.1371/journal.pone.0207475

Muñoz, M., Bozzi, R., García-Casco, J., Núñez, Y., Ribani, A., Franci, O., et al. (2019). Genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds assessed with a high density SNP chip. *Sci. Rep.* 9:13546. doi: 10.1038/s41598-019-49830-6

Ollivier, L., Alderson, L., Gandini, G. C., Foulley, J. L., Haley, C., Joosten, R., et al. (2005). An assessment of European pig diversity using molecular markers: partitioning of diversity among breeds. *Conserv. Genet.* 6, 729–741. doi: 10.1007/s10592-005-9032-6

Pérez-Enciso, M., Ferraz, A. L., Ojeda, A., and López-Béjar, M. (2009). Impact of breed and sex on porcine endocrine transcriptome: a bayesian biometrical analysis. *BMC Genomics* 10:89. doi: 10.1186/1471-2164-10-89

Pérez-Montarelo, D., Madsen, O., Alves, E., Rodríguez, M. C., Folch, J. M., Noguera, J. L., et al. (2014). Identification of genes regulating growth and fatness traits in pig through hypothalamic transcriptome analysis. *Physiol. Genomics* 46, 195–206. doi: 10.1152/physiolgenomics.00151.2013

Peters, J., Von Den Driesch, A., and Helmer, D. (2005). "The upper euphrates-tigris basin: cradle of agro-pastoralism," in *New Methods and the First Steps of Mammal Domestication: Proceedings of the 9th International Council of Archeozoology*, (Durham: Oxford), 96–123.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á, and Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front. Genet.* 4:98. doi: 10.3389/fgene.2013.00098

Porter, V. (2002). *Mason's World Dictionary of Livestock Breeds, Types and Varieties*. Wallingford: CABI Publishing, doi: 10.1079/9780851994307.0000

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1111/j.1471-8286.2007.01758.x

Puig-Oliveras, A., Ballester, M., Corominas, J., Revilla, M., Estellé, J., Fernández, A. I., et al. (2014). A Co- association network analysis of the genetic determination of Pig conformation, growth and fatness. *PLoS One* 9:e114862. doi: 10.1371/journal.pone.0114862

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

Purfield, D. C., McParland, S., Wall, E., and Berry, D. P. (2017). The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. *PLoS One* 12:e0176780. doi: 10.1371/journal.pone.0176780

R Core Team, (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramos, A. M., Crooijmans, R. P. M. A., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* 4:e6524. doi: 10.1371/journal.pone.0006524

Ratajczak, T., Cluning, C., and Ward, B. K. (2015). Steroid receptor-associated immunophilins: a gateway to steroid signalling. *Clin. Biochem. Rev.* 36, 31–52.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169

Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105, 767–779.

Ritzoffy, N. (1935). *Uzgoj Svinja (Pig breeding)*. Zagreb: Faculty of Agriculture in Zagreb.

Ropka-Molik, K., Bereta, A., Żukowski, K., Tyra, M., Piórkowska, K., Żak, G., et al. (2018). Screening for candidate genes related with histological microstructure, meat quality and carcass characteristic in pig based on RNA-seq data. *Asian-Australas. J. Anim. Sci.* 31, 1565–1574. doi: 10.5713/ajas.17.0714

Rousset, F. (2008). Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi: 10.1038/nature01140

Saura, M., Fernández, A., Rodríguez, M. C., Toro, M. A., Barragán, C., Fernández, A. I., et al. (2013). Genome-Wide estimates of coancestry and inbreeding in a closed herd of ancient Iberian Pigs. *PLoS One* 8:e78314. doi: 10.1371/journal.pone.0078314

Schäler, J., Krüger, B., Thaller, G., and Hinrichs, D. (2020). Comparison of ancestral, partial, and genomic inbreeding in a local pig breed to achieve genetic diversity. *Conserv. Genet. Res.* 12, 77–86. doi: 10.1007/s12686-018-1057-5

Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., et al. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5661–E5669. doi: 10.1073/pnas.1416991111

Silva, R. M. O., Stafuzza, N. B., de Oliveira Fragomeni, B., de Camargo, G. M. F., Ceacero, T. M., Cyrillo, J. N. D. S. G., et al. (2017). Genome-Wide association study for carcass traits in an experimental nelore cattle population. *PLoS One* 12:e0169860. doi: 10.1371/journal.pone.0169860

Šprem, N., Salajpal, K., Safner, T., Dikic, D., Juric, J., Curik, I., et al. (2014). Genetic analysis of hybridisation between domesticated endangered pig breeds and wild boar. *Livest. Sci.* 162, 1–4. doi: 10.1016/j.livsci.2013.12.010

Tang, K., Thornton, K. R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5:e171. doi: 10.1371/journal.pbio.0050171

Ulmansky, D. (1911). *Studie Uber Die Abstammung Des Šiška Schweines*. Vienna: Veterinärmedizinische Universität Wien.

Uremović, M. (2004). *Crna Slavonska Pasmina Svinja – Hrvatska Izvorna Pasmina*. Vukovar: Vukovarsko-srijemska županija.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072

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

Xia, J., Fan, H., Chang, T., Xu, L., Zhang, W., Song, Y., et al. (2017). Searching for new loci and candidate genes for economically important traits through gene-based association analysis of Simmental cattle. *Sci. Rep.* 7:42048. doi: 10.1038/srep42048

Yang, B., Cui, L., Perez-Enciso, M., Traspov, A., Crooijmans, R. P. M. A., Zinovieva, N., et al. (2017). Genome-wide SNP data unveils the globalization of domesticated pigs. *Genet. Sel. Evol.* 49:71. doi: 10.1186/s12711-017-0345-y

Yang, J., Li, W. R., Lv, F. H., He, S. G., Tian, S. L., Peng, W. F., et al. (2016). Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33, 2576–2592. doi: 10.1093/molbev/msw129

Zappaterra, M., Ros-Freixedes, R., Estany, J., and Davoli, R. (2018). Association study highlights the influence of ELOVL fatty acid elongase 6 gene region on

backfat fatty acid composition in large white pig breed. *Animal* 12, 2443–2452. doi: 10.1017/S1751731118000484

Zeder, M. (2017). "Out of the fertile crescent: the dispersal of domestic livestock through Europe and Africa," in *Human Dispersal and Species Movement: From Prehistory to the Present*, eds N. Boivin, R. Crassard, and M. Petraglia, (Cambridge: Cambridge University Press), 261–303. doi: 10.1017/9781316686942.012

Zhang, L., Orloff, M. S., Reber, S., Li, S., Zhao, Y., and Eng, C. (2013). cgaTOH: extended approach for identifying tracts of homozygosity. *PLoS One* 8:e57772. doi: 10.1371/journal.pone.0057772

Zhang, L., Zhou, X., Michal, J. J., Ding, B., Li, R., and Jiang, Z. (2014). Genome wide screening of candidate genes for improving piglet birth weight using high and low estimated breeding value populations. *Int. J. Biol. Sci.* 10, 236–244. doi: 10.7150/ijbs.7744

Zhu, J., Chen, C., Yang, B., Guo, Y., Ai, H., Ren, J., et al. (2015). A systems genetics study of swine illustrates mechanisms underlying human phenotypic traits. *BMC Genomics* 16:88. doi: 10.1186/s12864-015-1240-y

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 May 2020
doi: 10.3389/fgene.2020.00344

Genome Wide Assessment of Genetic Variation and Population Distinctiveness of the Pig Family in South Africa

*Nompilo Lucia Hlongwane[1,2], Khanyisile Hadebe[1], Pranisha Soma[3], Edgar Farai Dzomba[2] and Farai Catherine Muchadeyi[1]**

[1] Biotechnology Platform, Agricultural Research Council, Onderstepoort, South Africa, [2] Discipline of Genetics, School of Life Sciences, University of KwaZulu-Natal, Pietermartizburg, South Africa, [3] Animal Production Institute, Agricultural Research Council, Irene, South Africa

Genetic diversity is of great importance and a prerequisite for genetic improvement and conservation programs in pigs and other livestock populations. The present study provides a genome wide analysis of the genetic variability and population structure of pig populations from different production systems in South Africa relative to global populations. A total of 234 pigs sampled in South Africa and consisting of village (n = 91), commercial (n = 60), indigenous (n = 40), Asian (n = 5) and wild (n = 38) populations were genotyped using Porcine SNP60K BeadChip. In addition, 389 genotypes representing village and commercial pigs from America, Europe, and Asia were accessed from a previous study and used to compare population clustering and relationships of South African pigs with global populations. Moderate heterozygosity levels, ranging from 0.204 for Warthogs to 0.371 for village pigs sampled from Capricorn municipality in Eastern Cape province of South Africa were observed. Principal Component Analysis of the South African pigs resulted in four distinct clusters of (i) Duroc; (ii) Vietnamese; (iii) Bush pig and Warthog and (iv) a cluster with the rest of the commercial (SA Large White and Landrace), village, Wild Boar and indigenous breeds of Koelbroek and Windsnyer. The clustering demonstrated alignment with genetic similarities, geographic location and production systems. The PCA with the global populations also resulted in four clusters that where populated with (i) all the village populations, wild boars, SA indigenous and the large white and landraces; (ii) Durocs (iii) Chinese and Vietnamese pigs and (iv) Warthog and Bush pig. K = 10 (The number of population units) was the most probable ADMIXTURE based clustering, which grouped animals according to their populations with the exception of the village pigs that showed presence of admixture. AMOVA reported 19.92%–98.62% of the genetic variation to be within populations. Sub structuring was observed between South African commercial populations as well as between Indigenous and commercial breeds. Population pairwise F_{ST} analysis showed genetic differentiation ($P \leq 0.05$) between the village, commercial and wild populations. A per marker per population pairwise F_{ST} analysis revealed SNPs associated with QTLs for traits such as meat quality, cytoskeletal and muscle development, glucose

OPEN ACCESS

Edited by:
Maria Saura, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Spain

Reviewed by:
Kwan-Suk Kim, Chungbuk National University, South Korea
Manuel Vera, University of Santiago de Compostela, Spain

***Correspondence:**
Farai Catherine Muchadeyi
MuchadeyiF@arc.agric.za

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *07 September 2018*
Accepted: *23 March 2020*
Published: *07 May 2020*

Citation:
Hlongwane NL, Hadebe K, Soma P, Dzomba EF and Muchadeyi FC (2020) Genome Wide Assessment of Genetic Variation and Population Distinctiveness of the Pig Family in South Africa. Front. Genet. 11:344. doi: 10.3389/fgene.2020.00344

metabolism processes and growth factors between both domestic populations as well as between wild and domestic breeds. Overall, the study provided a baseline understanding of porcine diversity and an important foundation for porcine genomics of South African populations.

Keywords: pigs, diversity, population structure, genetic characterization, SNP60K

INTRODUCTION

Pigs were domesticated over 5,000 years ago, leading to the gradual and cumulative development of modern pig breeds with very distinctive phenotypes and production abilities (Zeder et al., 2006; Rothschild and Ruvinsky, 2010). Domesticated pig (*Sus Scrofa domesticus*) originated from the *Sus scrofa*, which is commonly known as the wild boar belonging to the Suidae family (Jones, 1998). This family includes species of wild pigs such as *Phacochoerus africanus* (Common warthog), *Potamochoerus larvatus* (Bush pig) and *Hylochoerus meinertzhageni* (Giant Forest hog) some that are indigenous to Africa (Jones, 1998). The Wild Boars are widely distributed covering areas such as Europe, Asia, and North Africa and were introduced as game species in all other continents including Africa (Jones, 1998; Scandura et al., 2011).

Pig breeds worldwide are either of well-defined ancestry or in certain instances crossbreds from populations of diverse origins (Amills et al., 2010). South African pig production consists of a commercial intensive sector with defined breeds and an extensive sector that is mainly associated with small-scale farmers in the rural areas. Village production system is characterized by non-descript populations raised under extensive low-input management. Commercial breeds such as the Large White, Landrace and Duroc have worldwide distribution in modern commercial farming systems including South Africa and are widely used (Amills et al., 2010). Indigenous breeds classified under *Sus indica* such as Kolbroek and Windsnyer are geographically restricted to Southern Africa (Nicholas, 1999). The Kolbroek, which is of Chinese origin, is speculated to have pigs that ended up in the hands of South African farmers when a sailing ship wrecked at the Cape Hangklip (Ramsay et al., 1994). Although the origin of the Windsnyer is unknown, there are observed similarities to Chinese breeds (Nicholas, 1999) thereby suggesting that it is of Chinese origin. Regardless of their origins and domestication routes, pig breeds in South Africa have become closed genetic pools restricted to specific farming systems and molded by artificial selection and possibly genetic drift (Amills et al., 2010). In addition to these domesticated breeds are the Warthog, Bush pig and Red River Hog wild pigs that are native to Africa and are found roaming in forests or in the zoos (Porter, 1993). The common Warthog (*Phacochoerus Africanus*) which was first discovered at Cape Verde, Senegal is one of the three species found in Africa. The Cape Warthog (*Phacochoerus aethiopicus*) is now extinct due to the rinderpest epizootic of the 1860s (Pallas, 1766; Gmelin, 1788; D'Huart and Grubb, 2003). Another Warthog (*Phacochoerus delamerei*) species was described in Somalia and later renamed *Phacochoerus aethiopicus delamerei* as it is similar to the Cape Warthog (Lönnberg, 1908, 1912; Roosenvelt and Heller, 1915). Muwanika et al. (2003) studied the phylogeography of the common Warthog in Africa and found three clades representing West, South and East African Warthogs. There is no enough evidence to support the origin of the Bush pig, which was assumed to have originated from Asia (White and Harris, 1977). There are recordings of the Bush pig in the Swellendam and Outeniqualand in the Western Cape provinces of South Africa (Rookmaaker, 1989). Hybrids between the domestic and Bush pigs have been recorded with the introduction of Bush pigs to South Africa being as far as 1400 years ago (Linnaeus, 1758; Mujibi et al., 2018). The existence of hybrids is a concern, as they could become asymptomatic carriers of diseases such African swine fever (Jori and Bastos, 2009).

Indigenous breeds are often geographically restricted and harbor unique genetic variants that may provide future breeds with the flexibility to change in response to product market preferences and production environments. While low-input and indigenous breeds may not compete with exotic breeds in terms of production performance, they are considered hosts to unique genetic diversity that should be protected as sources of variation. Local pigs are important because of their hardiness and ability to survive in extreme conditions (Taverner and Dunkin, 1996; Zadik, 2005). Most indigenous breeds are, however, threatened by small and fragmented flock sizes, which predispose them to lose genetic diversity as a result of genetic drift and indiscriminate crossbreeding with exotic germplasm that can lead to genetic erosion and the eradication of the local genetic pool. Globally, 35% of pig breeds are classified as at risk or already extinct (FAO, 2009) demonstrating the threat to local biodiversity.

Genomics have emerged as an effective tool for assessing diversity within and amongst populations. Swart et al. (2010) observed low differentiation among pig populations in Southern Africa using microsatellites. Heterozygosity levels ranged from 0.531 to 0.692 for commercial and indigenous breeds. The availability of the Porcine SNP60K BeadChip has opened new avenues of examining genetic diversity (Ramos et al., 2009) at a genome wide scale relative to that using microsatellite and other low-coverage markers. Mujibi et al. (2018) observed close clustering of Warthogs and Bush pigs using the Porcine SNP60K BeadChip. The Porcine SNP60K BeadChip has been used to infer on population structure and selection signatures in Chinese and European pig populations (Ai et al., 2013). Using this SNP panel in South African pig populations will provide comprehensive information on the genomic architecture of local, exotic and wild pig populations, which will guide future management and conservation. The objective of the

present study was to provide a large-scale analysis of the genetic diversity and structure of South African local pig populations using the Porcine SNP 60K BeadChip. The study investigated diversity of South African pigs relative to global populations of 389 pigs consisting of villages and out-group pigs from South America, Europe, United States, and China amongst other countries.

MATERIALS AND METHODS

Breeds/Populations Sampled

South African specimens were collected from a total of 234 samples from different production systems, representing village, intensively farmed populations in conservation units and free ranging populations. Village and non-descript pig populations were sampled from Alfred Nzo (ALN; n = 17) and Oliver Reginald Tambo (ORT; n = 22) districts in Eastern Cape province and Mopani (MOP; n = 27) and Capricorn (CAP; n = 25) districts in Limpopo province. Commercial pig breeds of Large White (LWT; n = 20), South African Landrace (SAL; n = 20) and Duroc (DUR; n = 20) were sampled from commercial farmers in Limpopo province. Indigenous populations Kolbroek (KOL; n = 20.) and Windsnyer (WIN; n = 20) were sampled from the Agricultural Research Council-Animal Production Institute in Pretoria, South Africa (**Table 1**). Vietnamese Potbelly breed (VIT; n = 5) was sampled from the Johannesburg Zoo and represents a breed that is endangered in Vietnam, its country of origin but has been raised in a conservation zoo in South Africa. European Wild Boar (n = 4), Warthogs (n = 31), and Bush pigs (n = 3) were sampled as representatives of the wild pig populations. The European Wild Boar and Bush pigs were sampled from the surrounding villages in the North-West whilst the Warthog samples were collected from geographically separated National Parks from North-West (n = 4), Eastern Cape (n = 3), and Limpopo (n = 24). The distribution of the sampled individuals is illustrated in **Figure 1**. Ear tissue samples were collected using the tissue sampling applicator gun while pliers were used to collect the hair samples according to standard procedures and ethical approval from ARC-Irene Animal Ethics committee (APIEC16/028).

TABLE 1 | Population category and sample size of the 13 pig populations.

Category	Population	Code	*N*
Village	Mopani	MOP	27
Village	Capricorn	CAP	25
Village	Oliver Reginald Tambo	ORT	22
Village	Alfred Nzo	ALN	17
Commercial	Large White	LWT	20
Commercial	SA Landrace	SAL	20
Commercial	Duroc	DUR	20
Indigenous	Kolbroek	KOL	20
Indigenous	Windsnyer Type	WIN	20
Asian	Vietnamese Potbelly	VIT	5
Wild	Wild Boar	WBO	4
Wild	Warthog	WAT	31
Wild	Bush Pig	BSP	3

Genotyping and Quality Control

DNA was extracted at the Agricultural Research Council-Biotechnology Platform from the ear tissue and hair samples using a commercially available Perkin Elmer Genomic DNA kit according to the manufacturer's protocol. DNA concentration was quantified using the Qubit® 2.0 Fluorometer. Gel electrophoresis (5%) was used to assess the quality and integrity of the DNA.

All 234 animals were genotyped using PorcineSNP60 v2 genotyping BeadChip (Illumina, United States) containing 62,163 SNPs with an average gap of 43.4 kb. Genotyping was done using the standard infinium assay at the ARC-Biotechnology Platform in South Africa. GenomeStudio version 2.0 (Illumina, United States) was used to process the genotype data, including raw data normalization, clustering and genotype calling. A final custom report was created to be able to generate a Plink *Ped* (Pedigree file) and *Map* (SNP panel file) for use in downstream analysis.

Golden Helix SNP Variation Suite (SVS) version 8.5 was used to update the SNPs marker file (Golden Helix Inc., 2016) based on the pig genome assembly (*Sus Scrofa* v10.2). Markers were then filtered to exclude SNPs located on the sex chromosomes. From this data set, Minor allele frequency (MAF) and deviation from Hardy–Weinberg equilibrium (HWE) were estimated per population for the 10 populations that excluded BSP, VIT, and WBO, which were left out due to small sample sizes. Additional quality control (QC) was also performed per population to remove SNPs with less than 85% call rate, MAF $<$ 0.02 and HWE $<$ 0.0001. The resultant filtered dataset was used to calculate observed (H_O), and expected (H_E) heterozygosities, inbreeding (F_{IS}) and effective population size (N_e).

Quality control was then performed overall population to remove SNPs with less than 85% call rate, MAF $<$ 0.02 and HWE $<$ 0.0001 and generate a dataset used for analysis of molecular variance (AMOVA) and F_{ST} analysis. Using this dataset, further QC filtered for SNPs in high LD (r^2 = 0.2) and closely related individual [Identity By Descent (IBD) $\geq$ 0.45] to produce a filtered dataset used for population structure analysis using ADMIXTURE and Principle Component Analysis (PCA).

Genetic Diversity Within Population

The MAF, H_E and H_O were calculated as measures of within population genetic variation using PLINK 1.07 (Purcell et al., 2007). In addition, inbreeding coefficient (F_{IS}) was calculated on Golden Helix SNP Variation Suite (SVS) version 8.5 (Golden Helix Inc., 2016). Effective population size (N_e) trends across generations were estimated based on a relationship between r^2 (expected LD), N_e and C (recombination rate). SNeP software (Version 1.1) tool was used based on the following formula suggested by Corbin et al. (2012) using the equation:

$$N_{T(t)} = \frac{1}{(4f(C_t))}\frac{1}{E[r^2_{adj}|C_t]} - \alpha.$$

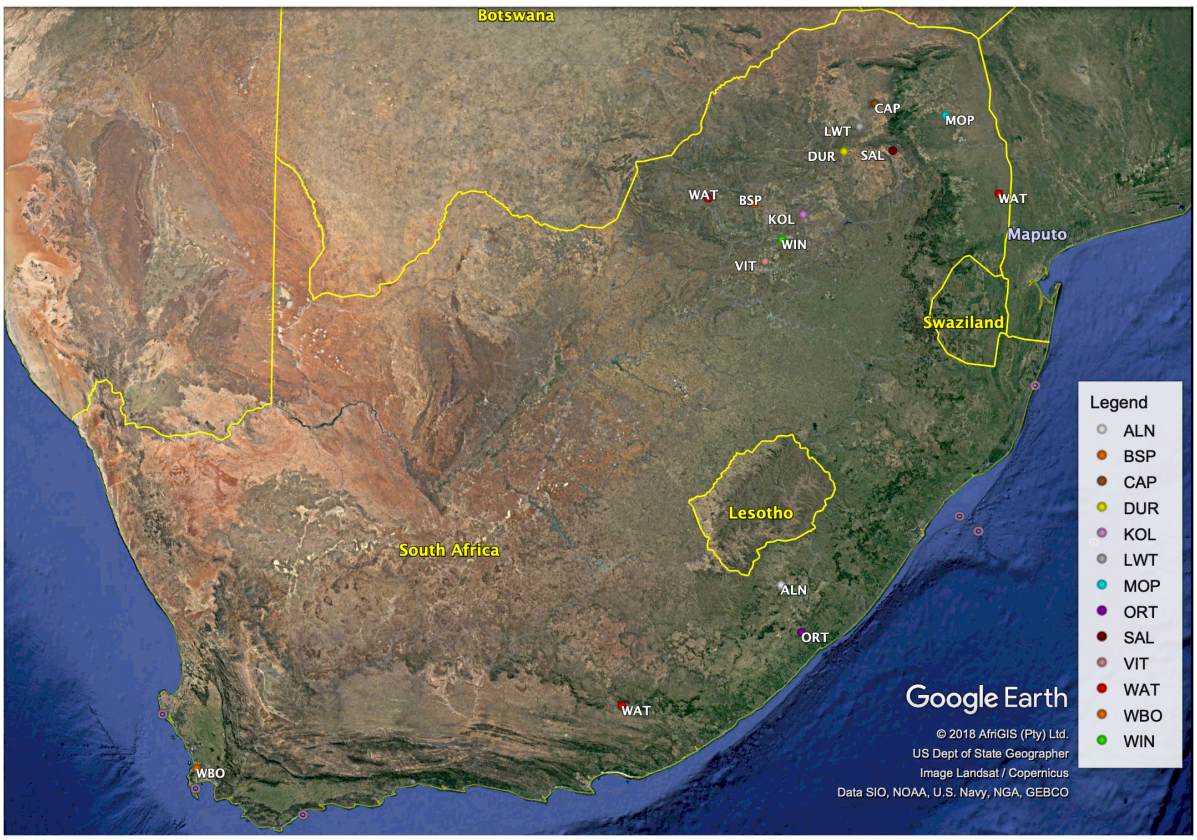


FIGURE 1 | Map showing geographic locations of the 13 pig populations in the present study.

where:

$N_{T(t)}$: Effective population size estimated t generations ago
C_t: Recombination rate t generations ago
$r_{2\text{adj}}$: Linkage disequilibrium estimation adjusted for sampling biasness
α: a constant.

The recombination rate was estimated by using the following formula proposed by Sved (1971):

$$f(c) = c\left[\frac{\left(1-\frac{c}{2}\right)}{(1-2)^2}\right].$$

The Bush pig, Vietnamese Potbelly and Wild Boar were excluded from the diversity within population analysis due to their small sample sizes. The few available samples were sampled from zoos and game reserves in the country where only few animals are often rescued and kept in conservation.

Population Differentiation and Structure

Analysis of Molecular Variance (AMOVA) was used to determine the genetic variance within populations (F_{IS}), among populations within group (F_{SC}) and among groups (F_{CT}) using ARLEQUIN v3.5 (Excoffier et al., 2005). The populations were categorized into villages, commercial, indigenous and wild populations and consisted of animals sampled in South Africa as well global populations from Burgos-Paz et al. (2013) which consisted of 389 genotypes of villages and out-group pigs from 24 countries of America (United States), South America (Mexico, Cuba, Guadeloupe, Guatemala, Costa Rica, Columbia, Ecuador, Peru, Brazil, Bolivia, Paraguay, Argentina, and Uruguay), Europe (Spain, Portugal, Italy, Poland, Hungary, Tunisia, Denmark, Holland, United Kingdom) and China. Variance components were also estimated for groups consisting of different categories, i.e., village and indigenous; indigenous and commercial; South African village and global villages; South African commercial and global commercial etc.

Principal Component Analysis (PCA) using SVS version 8.5 (Golden Helix Inc., 2016) and the eigenvector method was used to determine population clustering. ADMIXTURE version 1.20 (Alexander and Lange, 2011) was used to detect the most likely clusters (K) for the population. ADMIXTURE was run from $K = 2$ to $K = 15$. The number of potential genetic clusters (K) was tested from 1–15 to reassign each sample to its population of origin. The optimum K-value was that with the lowest cross-validation error value. Initially, all the 13 populations sampled from South Africa were included in the population structure analysis. After this the South African data set was merged to Porcine SNP60K genotype data from Burgos-Paz et al. (2013) described above.

Population pairwise F_{ST} values were estimated according to the formula of Weir and Cockerham (1984) implemented in the Golden Helix SNP Variation Suite (SVS) version 8.5 (Golden Helix Inc., 2016). Based on population pairwise F_{ST} values, PCA and ADMIXTURE based clustering, F_{ST} analysis per marker was estimated between pairs of highly differentiated populations of the village populations, indigenous populations and commercial breeds as well as amongst highly differentiated commercial breeds and wild populations. To reduce noise, an F_{ST} averaged smooth value was used to identify genomic regions differentiating pairs of populations. Manhattan plots of per marker F_{ST} values between pairs of populations were plotted against chromosomal coordinates using the porcine assembly (*Sus Scrofa* 10.2). Highly differentiating SNPs ($F_{ST} \geq 0.8$) were subsampled and genes associated with these SNPs searched using genome browse including their associations with known QTLs in the pig genome based on the *Sus Scrofa* 10.2 on Ensembl[1].

RESULTS

Genotypes and Quality Control

The percentage of polymorphic and number of SNPs (N_{SNP}) remaining after QC per population and overall is presented in **Table 2**. Two hundred and eleven individuals with a genotyping rate of 85% remained after QC. Windsnyer pigs had the highest percentage of informative markers (95%) after QC, whilst Warthog had the lowest at 82%. About 31,705 SNPs were removed leaving 30,458 polymorphic SNPs of the loci distributed over 18 autosomal chromosomes, which were used for AMOVA and F_{ST} analysis. After LD and IBD pruning, 23,345 SNPs and 176 individuals were used for the population structure analysis.

Genetic Diversity Across Populations

Genetic diversity parameters among the 10 populations are summarized in **Table 2**. Warthog pigs had the lowest H_O (0.188 ± 0.155) and Windsnyer the highest (0.385 ± 0.171). Expected heterozygosity values ranged from 0.204 ± 0.151 from Warthog to 0.371 ± 0.126 for Capricorn. The highest inbreeding coefficient (F_{IS}) was for Warthog at 0.398 ± 0.475 while the Duroc had the lowest and slightly negative value of -0.067 ± 0.153. F_{IS} values were positive for all village populations as well as Warthog suggesting some level of inbreeding within these populations. MAF was the highest in village population from Capricorn (0.264 ± 0.147) and the least in Warthog pigs (0.076 ± 0.109).

Effective Population Size

Figure 2 shows trends in effective population size across all of the studied populations. The Warthog was excluded in this analysis because the number of polymorphic SNPs was not enough to generate results. Effective population size values are presented in **Supplementary Table S1**. There was a general decline in N_e across all the populations across generations. The indigenous and commercial populations had higher effective population size compared to the village populations. The Kolbroek had the lowest effective population size 12 generations prior.

AMOVA

Genetic differentiation between populations is presented in **Supplementary Table S2**. The major proportion of the genetic variance was attributed to variation within South African populations with F_{IS} values ranging from 76.41 to 98.62%. Diversity within populations (F_{IS}) in village populations from this study and those from Burgos-Paz et al. (2013) was 35.52% while variation among groups (F_{CT}) was 62.35%. Diversity of South African commercial pigs was 76.41% within populations, 18.17% among populations within group and 5.42% among groups. When including the commercial breeds from Burgos-Paz et al. (2013), the diversity parameters changed to F_{IS} = 30.97%, F_{SC} = 8.31% and F_{CT} = 60.72%. High F_{CT} (>60%) were observed in the category consisting of South African indigenous and Chinese indigenous (F_{CT} = 70.08%) as well as that consisting of the South African Wild Boar and the worldwide Wild Boar (F_{CT} = 73.58%).

[1] www.ensembl.org

TABLE 2 | Summary of the genetic diversity measures across South African Pig populations.

POP	*N*	$\%_{SNP}$	MAF ± SD	N_{SNP}	H_O ± SD	H_E ± SD	F_{IS} ± SD	*P*-value
MOP	27	92	0.262 ± 0.149	52,925	0.299 ± 0.129	0.369 ± 0.131	0.198 ± 0.134	0.495
CAP	24	94	0.264 ± 0.147	54,078	0.332 ± 0.140	0.371 ± 0.126	0.117 ± 0.155	0.582
ORT	22	93	0.259 ± 0.153	52,238	0.315 ± 0.145	0.370 ± 0.130	0.163 ± 0.113	0.553
ALN	15	94	0.238 ± 0.157	53,580	0.336 ± 0.160	0.359 ± 0.134	0.056 ± 0.168	0.695
LWT	18	93	0.227 ± 0.161	49,773	0.358 ± 0.177	0.348 ± 0.144	0.023 ± 0.009	0.721
SAL	19	94	0.221 ± 0.162	49,191	0.372 ± 0.186	0.345 ± 0.144	0.052 ± 0.085	0.704
DUR	19	94	0.177 ± 0.168	40,632	0.359 ± 0.182	0.337 ± 0.147	0.067 ± 0.153	0.764
KOL	20	94	0.173 ± 0.167	39,560	0.364 ± 0.182	0.339 ± 0.144	0.051 ± 0.087	0.727
WIN	19	95	0.220 ± 0.164	47,402	0.385 ± 0.171	0.360 ± 0.134	0.056 ± 0.158	0.733
WAT	28	82	0.076 ± 0.109	3,967	0.188 ± 0.155	0.204 ± 0.151	0.398 ± 0.475	0.710

$\%_{SNP}$ used to calculate MAF analysis; N_{SNP}, the number of SNPs in the subset 62,163 SNP; H_O, observed heterozygosity; H_E, expected heterozygosity; SD, standard deviation; F_{IS}, inbreeding co-efficient; MAF, minor allele frequency, $P < 0.05$.

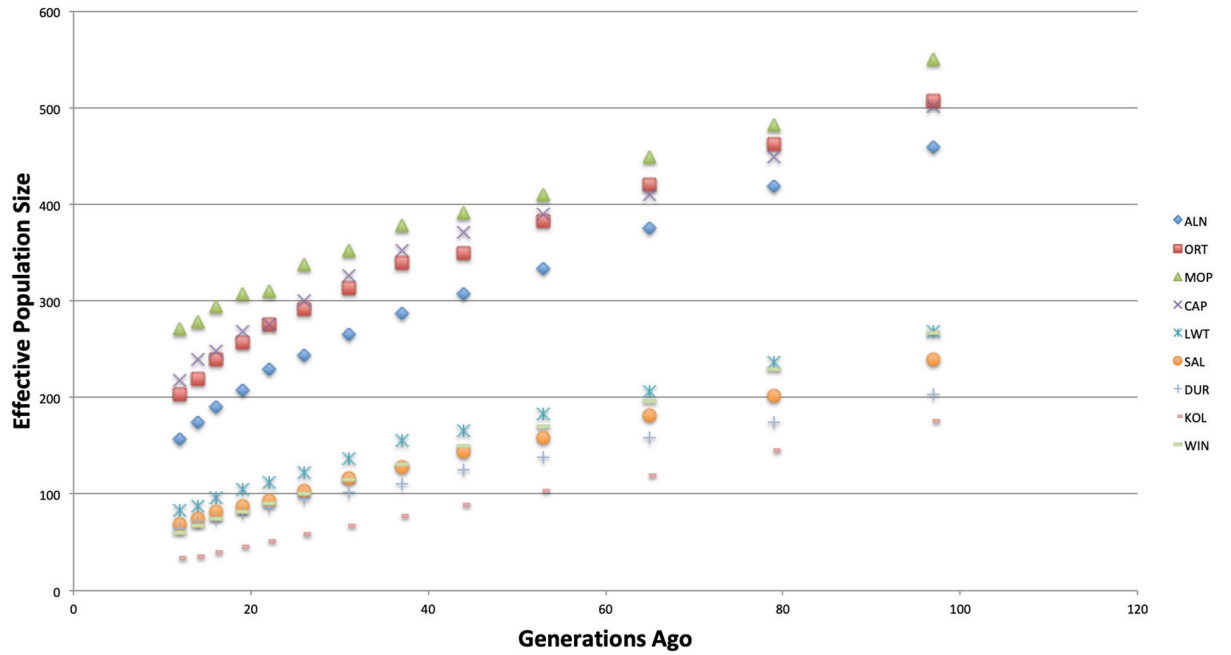


FIGURE 2 | Average effective population size plotted against generation in the past.

Population Structure

Principal component one (PC1) and principal component two (PC2) explained approximately 30.7% and 11.8% of the total variation, respectively. The PCA of South African breeds yielded four main genetic clusters (**Figure 3**). The Duroc clearly separated from the Large White and South African Landrace that clustered together with the wild boar and village populations. The Warthog and the Bush pig clustered together as a third cluster whilst the fourth cluster consisted of Vietnamese potbelly sampled from the zoo. The PCA analysis using South African samples and those from Burgos-Paz et al. (2013) demonstrated the same clustering with all the village pigs grouping together with the Large White and Landraces separated from clusters of (i) Warthog and Bush pig, (ii) Chinese and Vietnamese breeds and (iii) Duroc (**Figure 4**).

Genetic structure of the South African breeds was further investigated using ADMIXTURE. The results presented in **Figure 5** show the Warthog and Bush pig populations clustering together and clearly separated from the rest of the other populations at $K = 2$. Duroc separated from the rest of the populations at $K = 3$ followed by Vietnamese at $K = 4$. $K = 4$ clustered animals in the same way observed with PCA based clustering. Beyond $K = 8$, the genetic clusters of the commercial, indigenous, Asian and wild breeds are maintained whilst the added K is distributed within the village populations. $K = 10$ which was the optimal K (**Supplementary Figure S1**) with lowest CV (0.551) resulted in the eight distinct genetic clusters of commercial, indigenous, Asian and wild breeds plus highly admixed clusters consisting of all village pig populations from Limpopo and Eastern Cape provinces of South Africa.

Population Differentiation

Population pairwise F_{ST} values are shown in **Table 3**. Low F_{ST} were observed between village populations with values ranging from 0.022–0.060 ($P < 0.05$) within South Africa and in global populations. The highest differentiation was found between Warthog and Duroc at $F_{ST} = 0.481$. Warthog and Kolbroek pigs showed the high differentiation at 0.468. All other populations had F_{ST} values above 0.282. The extent of differentiation between Warthog and all the other populations was high ranging from 0.312 (Warthog and Creole from Columbia) to 0.589 (Warthog and Vietnamese). Highest F_{ST} observed was between Vietnamese and Bush pig populations at 0.700 (**Supplementary Table S3**).

Per Marker Pairwise F_{ST}, SNP Annotation and Association With Porcine QTLs

Per population, per marker pairwise F_{ST} values were computed for highly differentiated populations and are illustrated in **Table 4**, **Supplementary Figure S2**. SNPs. High F_{ST} values ($\geq$0.8) where considered breed differentiating and the associated SNPs were functionally annotated for genes within a 1 MB region. Fixed SNPs ($F_{ST} = 1.0$) where observed on chromosome 9 between Duroc and Warthog, on chromosome 12 between Koelbroek and Warthog and on chromosome 18 between Windsnyer and Warthog. For all the pairwise comparisons, 281 SNPs ($F_{ST} \geq 0.8$) were detected (**Supplementary Figure S2**) with only 123 candidate genes within 1 MB of those SNPs. Pairwise comparison of village pigs from Alfred NZO, South Africa and Warthog yielded genes related to acute heat stress (*RPL18*) and inflammatory response (*IL17B* and *ARHGAP23*) as illustrated

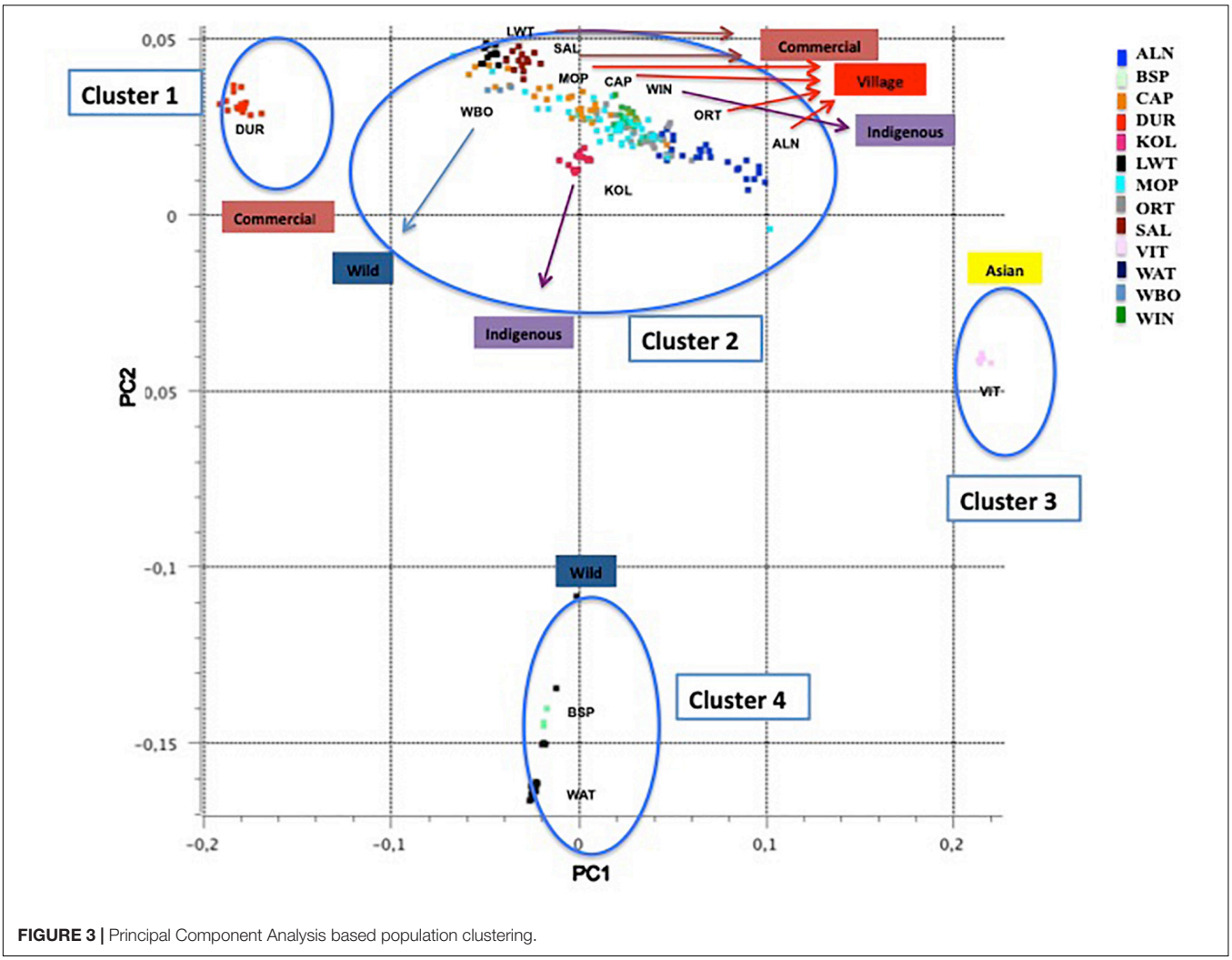


FIGURE 3 | Principal Component Analysis based population clustering.

in **Table 4** and **Supplementary Figure S2a**. Gene *ADGRB3* was in close proximity of SNPs *rs81353971, rs81353988, rs81353991, rs81297001,* and *rs81333295* that were of significant between Duroc and Warthog. Inflammatory response genes such as *ARHGAP23* were associated with the significant SNPs observed between Koelbroek, Large White and Windsnyer populations. For reproduction traits, genes *CD28*, *TCP11L2*, *TLK1*, *ATPB2, GPR137C, ZNF609*, *ARHGAP22, EPSTI1, GPR63, TCTE3, PTP4A2, ZSCAN20, CLU,* and *CACNA2D3* were observed within 14 significant SNPs on chromosomes 1, 2, 5, 6, 11, 14, and 15. Genes that had association with meat traits such as *DLX1*, *BRPF*1, *CLPTM1*, *FANCD2*, *SEC13*, *FHL3*, *FSTL5*, *CEP135, EXOC1, FOXO1, ASTN2, MYO18B, PLXNA1, DNAH2, HECTD2, TMEM39B, TXLNA, CSMD2, COL16A1, SCARA3, ZFAND3,* and *PTPRD* were also reported. Comparison with indigenous pigs showed genes that were associated with mastitis resistance (*ARHGAP39*, *ARPC4, PHC2,* and *BCL2L15*) and hair follicle development (*FOXN1*). A total of eight SNPs associated with growth traits (*ADGRB3*, *TSPAN,* and *ZFAND3*) were detected. *PTPN3* gene associated with immune response was observed between indigenous and Wild Boar. Wild Boar and Duroc comparison resulted in genes associated with adaptation (*HDAC1* and *GNAI3*).

DISCUSSION

The Porcine SNP60K BeadChip was developed in 2009 (Ramos et al., 2009) and has been used to analyze genetic diversity and population structure in several pig populations (Ai et al., 2013; Burgos-Paz et al., 2013; Yang et al., 2017; Mujibi et al., 2018). This is the first report using the Porcine SNP60K BeadChip to explore diversity of domestic and wild pig populations covering the commercial, village, wild and conserved pigs farmed and reared in Africa. Pigs are possibly known to have reached Sub-Saharan Africa through the Nile corridor and later dispersed to the West-Central Africa (Blench, 2000). There are 541 pig breeds worldwide (Rischkowsky and Pilling, 2007) but the dominating commercial breeds in the pork industry are the Large White, Landrace, Duroc, Hampshire, Berkshire and Piétrain (Rothschild and Ruvinsky, 2010). The source of the improved breeds found in Southern Africa is believed to be the European settlers in

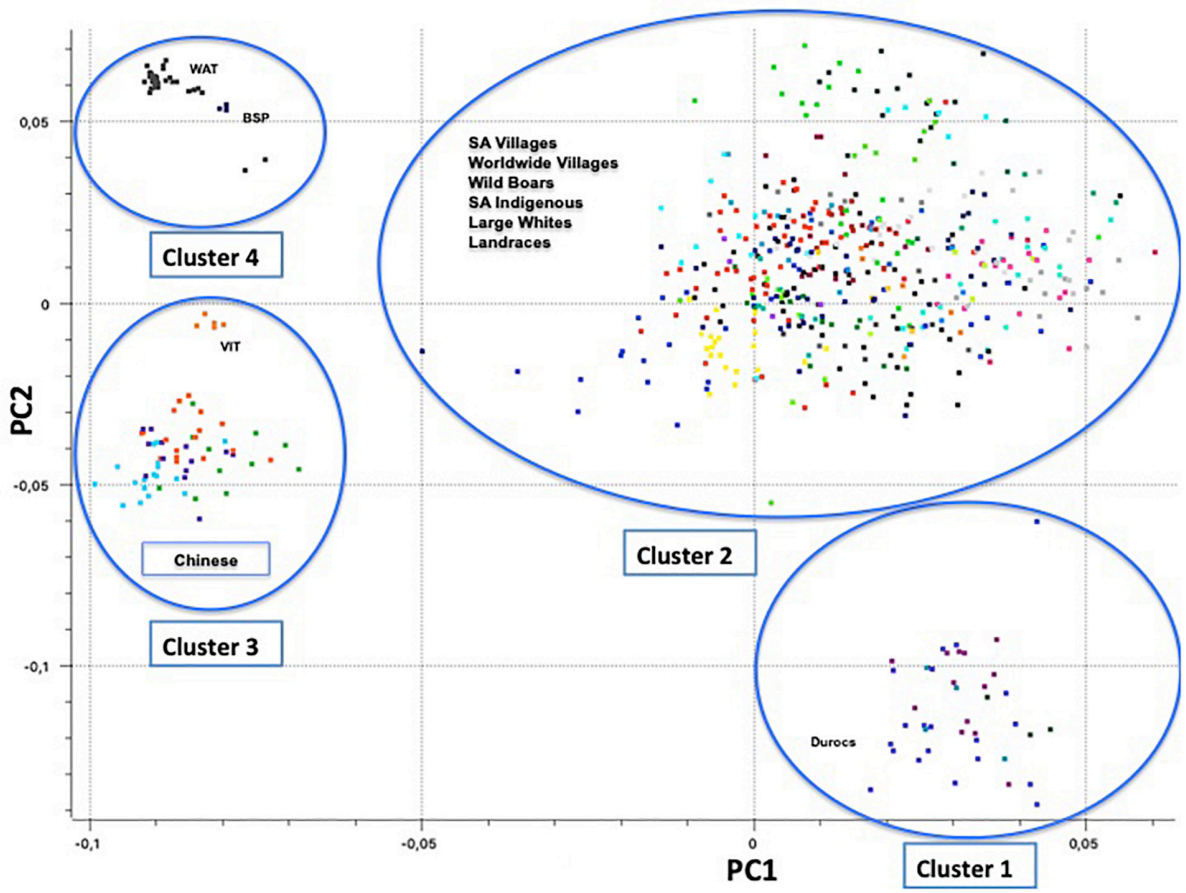


FIGURE 4 | Principal Component Analysis based population clustering including Burgos-Paz et al. (2013) genotypes (22,430 SNPs).

1600s (Krige, 1950; Blench and MacDonald, 2000; Swart et al., 2010). This was when Jan van Riebeeck brought some pigs to the Cape of Good Hope (Naude and Visser, 1994). The Large White, South African Landrace and the Duroc are the breeds mostly found and used in the commercial sector while the Kolbroek and Windsnyer are considered as indigenous and are mostly found in rural areas (Kem, 1993; Ramsay et al., 2000). The Vietnamese, Bush pig and Wild Boar populations constitute a small component of the genetic pool of pigs in the country often restricted to the game reserves and zoos.

The Porcine SNP60K BeadChip was designed using genomic resources from Western pig genomes (Ramos et al., 2009) and hence the number of SNPs after QC for the commercial population was higher (**Table 2**). The village populations had a higher number of polymorphic SNPs and moderate-high MAF compared to that of commercial pigs. Non-descript livestock populations including pigs are often observed to be highly diverse probably due to open mating systems and gene flow between populations. In South Africa similar observations of highly diverse and polymorphic populations were observed in village chicken populations (Khanyile et al., 2015), cattle (Makina et al., 2014), and village goats (Mdladla et al., 2016). The Warthog and other indigenous pigs were observed to be the least polymorphic and diverse which could be attributed to ascertainment bias as the Kolbroek, Windsnyer, Vietnamese Potbelly, Warthog and Bush pigs were not used in the development of the Porcine SNP60K BeadChip. Overall, the porcine SNP panel showed moderate MAF for the village, commercial and indigenous purebred pig populations such as the Windsnyer implying utility of the chip in the prevalent farmed pig populations of South Africa.

A study conducted by Swart et al. (2010) using microsatellite markers in various Southern African pig breeds revealed higher levels of diversity within population than was observed in this study for the same breeds (**Table 2**). High heterozygosity levels (0.61–0.75) were also reported by Halimani et al. (2012). In contrast to Swart et al. (2010) the Large White had the lowest diversity (H_o = 0.358) compared to the South African Landrace (H_o = 0.372) and other breeds of the Duroc and Kolbroek. It must be noted that these previous studies used microsatellite markers that are highly polymorphic markers and cannot be compared to SNPs that are biallelic in nature. High gene diversity is therefore expected in microsatellites markers. However, results on genetic diversity from this study were comparable to other studies that used the Porcine SNP60K BeadChip in Chinese and Western pig populations (Ai et al., 2013).

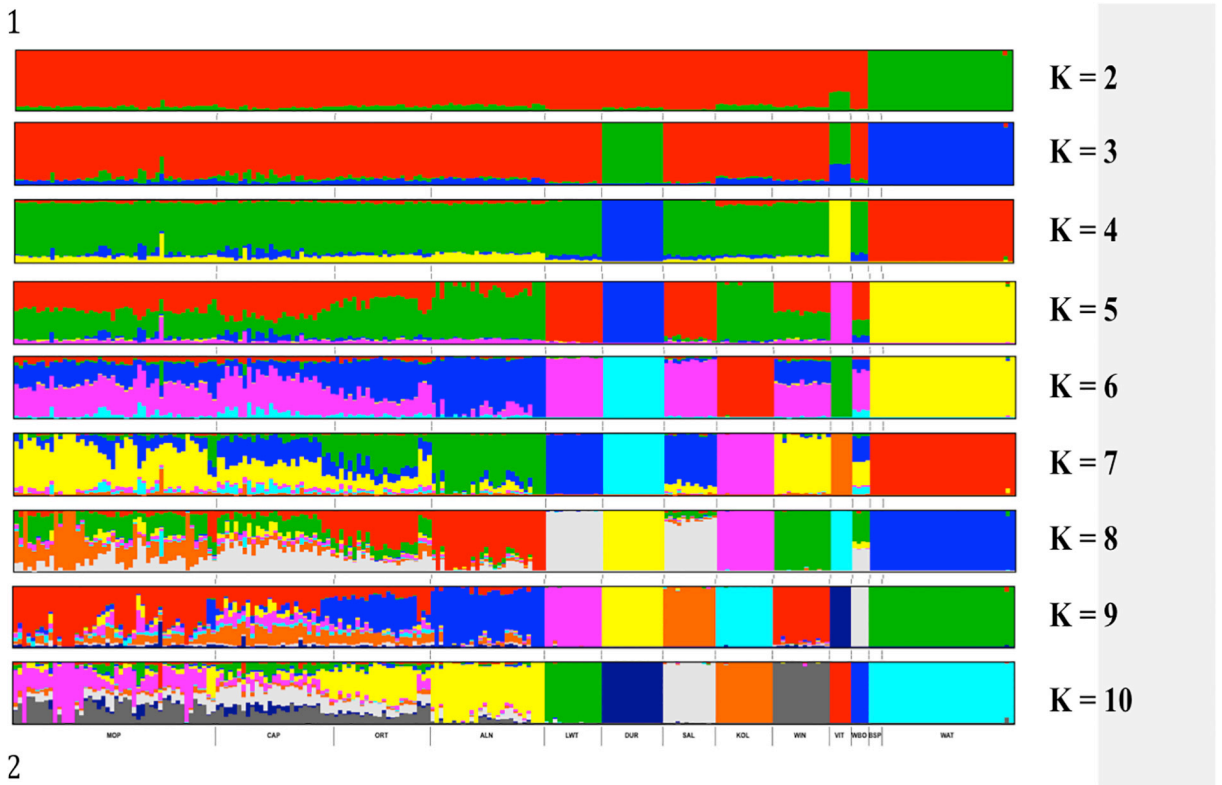


FIGURE 5 | ADMIXTURE based clustering $K = 2 - K = 10$. Each individual is represented by a single column divided into K colored segments, where K is the number of clusters assumed with lengths proportional to each of the K inferred cluster.

TABLE 3 | Pairwise genetic differentiation (F_{ST} values) between 10 pig populations.

		MOP	CAP	ORT	ALN	LWT	DUR	SAL	KOL	WIN	WAT
Villages	**MOP**										
	CAP	0.022*									
	ORT	0.031*	0.026*								
	ALN	0.059	0.060	0.040*							
Commercial	**LWT**	0.091	0.073	0.096	0.130						
	DUR	0.134	0.126	0.143	0.174	0.183					
	SAL	0.094	0.073	0.099	0.132	0.120	0.194				
Indigenous	**KOL**	0.120	0.116	0.129	0.162	0.189	0.237	0.194			
	WIN	0.061	0.064	0.077	0.106	0.143	0.189	0.144	0.173		
Wild	**WAT**	0.282	0.306	0.314	0.350	0.433	0.481	0.435	0.468	0.410	

*Significant levels: *P < 0.05.*

The heterozygosity values for the indigenous pigs were relatively similar to those of the commercial pigs (**Table 2**). A lower diversity was expected for the commercial pigs as they are under selection while the indigenous pigs are known to be rich reservoirs of distinct alleles, coupled with presence of gene flow (Amills et al., 2012). However, the indigenous pig populations are also of very small flock sizes and often fragmented and restricted to specific farming communities and conservation units hence diversity was low. Small and fragmented populations and the possibility of natural selection due to disease and unfavorable climatic conditions could explain the genetic diversity observed in the village populations. The high inbreeding levels observed in the Warthog populations might have been promoted by its family structuring where pigs are organized into fragmented breeding and social units (**Table 2**). Somers et al. (1995) noted that a group of Warthogs consist of about 40% of adults with changes seasonally. The number of mature individuals is estimated to be between 2000 and 5000 in the Kruger National Park (Ferreira et al., 2013). The geographical separation of the three national parks from which the warthogs were sampled, could have created small and fragmented subpopulations leading to escalated F_{IS} values

TABLE 4 | Most significant SNPs detected with F_{ST} analysis and the associated genes.

Population	SNP	Chr	Position	Genes	Function
ALN and WAT	rs81355030	1	84,376,735	*RPL18*	Acute heat stress (Newton et al., 2012)
	rs81367521	2	150,546,025	*IL17B*	Embryonic development, tissue regeneration and inflammation (Bie et al., 2017)
	rs81285672	12	23,638,629	*ARHGAP23*	Inflammatory response (Liu, 2015)
DUR and	rs81353971	1	49,024,494	*ADGRB3*	Growth traits (Emrani et al., 2017)
WAT	rs81353988	1	49,350,539	*ADGRB3*	
	rs81353991	1	49,392,902	*ADGRB3*	
	rs81297001	1	49,458,254	*ADGRB3*	
	rs81333295	1	49,592,586	*ADGRB3*	
	rs80946298	13	33,531,504	*DOCK3*	Induces axonal growth (Kimura et al., 2016)
	rs81444796	13	33,481,604	*DOCK3*	
	rs81478683	13	34,024,632	*IQCF3*	Conjunctival UV to auto fluorescence (Yazar et al., 2015)
	rs81478482	13	34,117,528	*ACY1*	Amino acid and heat shock protein (Martínez-Montemayor et al., 2008)
	rs81454214	15	107,134,695	*CD28*	Endometrial gene expression (Gu et al., 2014)
KOL and WAT	rs81341610	3	4,508,681	*LOC102160627*	Uncharacterized
	rs80993200	4	234,605	*ARHGAP39*	Milk production related and mastitis resistance (Wang et al., 2015)
	rs80851822	5	13,913,761	*POLR3B*	Residual feed intake (Gondret et al., 2017)
	rs80873063	5	13,940,475	*TCP11L2*	Regulated in small atretic follicles for healthy follicles (Hatzirodos et al., 2014a)
	rs80999600	5	66,998,856	*TSPAN9*	ADG (Fontanesi et al., 2014)
	rs80929588	5	67,092,749	*TSPAN9*	
	rs80883075	5	67,132,255	*TEAD4*	Regulation in organ size control and cell proliferation (Frankenberg et al., 2016)
	rs81385003	5	67,297,728	*ITFG2*	Disease resistance (Moioli et al., 2016)
	rs81285672	12	23,638,629	*ARHGAP23*	Inflammatory response (Liu, 2015)
	rs81325261	12	44,771,203	*FOXN1*	Regulation of hair follicle development (Song et al., 2017)
	rs80801871	13	33,170,033	*DOCK3*	Induces axonal growth (Kimura et al., 2016)
	rs80802886	13	33,202,454	*DOCK3*	
	rs81444784	13	33,306,071	*DOCK3*	
	rs81444796	13	33,481,604	*DOCK3*	
	rs80946298	13	33,531,504	*DOCK3*	
	rs81478683	13	34,024,632	*IQCF3*	Conjuctival UV to auto fluorescence (Yazar et al., 2015)
	rs335091311	15	148,461	*STAM2*	Residual feed intake (Gondret et al., 2017)
	rs80852223	15	77,232,829	*TLK1*	Decrease expression in the endometrium (Gray et al., 2006)
	rs80999734	15	77,318,065	*TLK1*	
	rs81453662	15	78,190,260	*DLX1*	Muscling and meat availability (Li et al., 2010)
LWT and WAT	rs81349766	1	182,224,202	*GPR137C*	Litter size (Sosa-Madrid et al., 2018)
	rs81296498	1	182,722,677	*DDHD1*	Lipid metabolism (Parker Gaddis et al., 2018)
	rs81349773	1	182,756,343	*DDHD1*	
	rs332395415	1	246,195,557	*ABCA1*	Mediates the transport of excess cholesterol (Schwartz et al., 2000)
	rs321979518	1	246,199,966	*ABCA1*	
	rs81383185	5	21,606,108	*RNF41*	Lipid rafts in immune signalling (McGraw and List, 2017)
	rs80820161	5	21,745,636	*STAT2*	Milk production (Salehi et al., 2015)
	rs80894897	5	21,727,701	*PAN2*	Fat yield (Suchocki et al., 2016)
	rs80940129	5	21,970,939	*BAZ2A*	Nutrition related (Cornelis and Hu, 2013)
	rs325229936	5	22,338,939	*MYO1A*	Coat color and pigmentation (Gutiérrez-Gil et al., 2007)
	rs81285672	12	23,6386,29	*ARHGAP23*	Inflammatory response (Liu, 2015)
	rs80854565	14	89,185,576	*ARHGAP22*	Fertility (Browett et al., 2018)
	rs80833618	14	89,227,581	*ARHGAP22*	
	rs80957034	14	89,255,703	*ARHGAP22*	
	rs80962102	14	89,309,115	*ARHGAP22*	
SAL and WAT	rs81395957	6	51,328,753	*NECTIN2*	Cell recognition and adhesion (Wang et al., 2010)
	rs81395929	6	51,427,663	*CLPTM1*	Marbling score (Lim et al., 2013)
WIN and WAT	rs81381252	4	65,339	*ZNF609*	Fertility (Hatzirodos et al., 2014b)
	rs81285672	12	23,638,629	*ARHGAP23*	Inflammatory response (Liu, 2015)

(Continued)

TABLE 4 | Continued

Population	SNP	Chr	Position	Genes	Function
	rs81325261	12	44,771,203	*FOXN1*	Regulation of hair follicle development (Song et al., 2017)
	rs331955329	13	66,004,327	*MTMR14*	Reduced with age accelerates skeletal muscle aging (Romero-Suarez et al., 2010)
	rs80971430	13	66,026,240	*BRPF1*	Intramuscular fatty acid (Puig-Oliveras et al., 2016)
	rs80945527	13	66,104,857	*ARPC4*	Mastitis resistance (Grossi et al., 2014)
	rs80885182	13	66,270,725	*FANCD2*	Muscle weight (Lionikas et al., 2010)
	rs45430493	13	66,515,894	*SEC13*	Muscle weight (Lionikas et al., 2012)
	rs81248260	13	66,583,753	*ATPB2*	Heat stress on reproductive performance (Dash et al., 2016)
	rs81446451	13	66,668,301	*ATPB2*	
	rs81446497	13	66,691,206	*ATPB2*	
	rs81446475	13	66,725,741	*ATPB2*	
	rs81446484	13	66,777,686	*ATPB2*	
	rs81478601	13	66,795,578	*ATPB2*	
IND and DUR	rs80866460	4	106,698,421	*PTPN22*	Immune response (Lamsyah et al., 2009)
	rs81413279	9	79,010,742	*NXPH1*	DMI (Olivieri et al., 2016)
	rs81413279	9	79,010,742	*ABCB5*	Immune function (Lee et al., 2017)
	rs81306790	6	89,661,963	*PHC2*	Mastitis (Chen et al., 2015)
	rs80854994	4	106,719,032	*PTPN22*	Immune response (Lamsyah et al., 2009)
	rs80854994	4	106,719,032	*BCL2L15*	Mastitis (Chen et al., 2015)
Villages and DUR	rs81282695	6	94,442,844	*POU3F1*	Neurobehavioral functioning (Eusebi et al., 2018)
	rs81282695	6	94442844	*FHL3*	Carcass traits (Zuo et al., 2004, 2007)
Villages and KOL	rs81430450	11	24,063,007	*DNAJC15*	Feeding efficiency (Reyer et al., 2017a)
	rs81430450	11	24,063,007	*EPSTI1*	Fertility traits (Gaddis et al., 2016), fat deposition (Zhang et al., 2018)
SAL&LWT and IND	rs81232179	8	51,070,662	*FSTL5*	Meat quality (Ryu and Lee, 2016); skeletal muscle (Novianti et al., 2010)
	rs45431508	8	69,912,174	*CXCL8*	Pig disease (Wang et al., 2019)
	rs81400554	8	55,181,102	*CEP135*	Intramuscular fat (Hamill et al., 2012); milk production (Rui et al., 2013)
	rs81400554	8	55,181,102	*EXOC1*	Marbling score (Wu et al., 2016)
	rs81400740	8	63,119,376	*EPHA5*	Feed efficiency (Reyer et al., 2017b)
	rs81400500	8	52,213,568	*NPY5R*	Feed efficiency and fat deposition (Chen et al., 2018)
	rs81400500	8	52,213,568	*NPY1R*	Feed efficiency and fat deposition (Chen et al., 2018)
	rs81302014	8	69,950,857	*RASSF6*	Body conformation (Fang and Pausch, 2019)
	rs80904678	11	15,274,089	*FOXO1*	Meat quality and carcass traits (Ropka-Molik et al., 2018)
	rs81400500	8	52,213,568	*SLC7A11*	Feed efficiency (Vigors et al., 2016)
	rs81300083	9	78,940,661	*NXPH1*	DMI (Olivieri et al., 2016)
IND and VIT	rs81350922	1	257,096,974	*ASTN2*	Carcass weight in cattle (Júnior et al., 2016)
	rs80970078	14	43,524,181	*MYO18B*	Meat quality and carcass traits (Ropka-Molik et al., 2018)
	DRGA0006738	6	117,857,953	*NOL4*	Fatness (Li et al., 2011)
	rs80860919	1	64,018,444	*GPR63*	Fertility traits (Moran et al., 2017)
	rs80921694	13	73,023,057	*PLXNA1*	Meat quality (Martínez-Montes et al., 2016)
	rs81327396	12	53,063,765	*DNAH2*	Intramuscular fat (Luo et al., 2012); carcass weight (Kang et al., 2013)
Villages and WBO	rs81244815	2	50,167,007	*SWAP70*	Disease resistance (Ma et al., 2011; Zhang et al., 2018)
	rs81244815	2	50,167,007	*SBF2*	Fertility (Zhang et al., 2014); immune function (Ibeagha-Awemu et al., 2016)
	rs81401075	8	73,841,435	*FRAS1*	Sow reproductive traits (Fischer et al., 2015), feed efficiency (Messad et al., 2019)
	rs81401075	8	73,841,435	*NPY2R*	Obesity (Siddiq et al., 2007; Hunt et al., 2011)
Villages and VIT	INRA0003181	1	95,198,598	*SLC14A2*	Conformation traits (Le et al., 2017)
	rs81332040	6	45,777,816	*ZNF382*	Conformation traits (Le et al., 2017)
	INRA0045852	14	10,3086,988	*HECTD2*	Fat and meat quality traits (Piórkowska et al., 2018)
	rs80980839	4	93,722,493	*RHBG*	Ammonia transporter (Xiang et al., 2016)
	rs80971176	5	49,876,132	*SOX5*	Ear morphology (Edea et al., 2017)

(Continued)

TABLE 4 | Continued

Population	SNP	Chr	Position	Genes	Function
WBO and DUR	rs80837120	1	565,627	*TCTE3*	Involved in spermatogenesis (Du et al., 2016)
	rs81389959	6	88,334,239	*PTP4A2*	Reproductive traits (Verardo et al., 2016); intramuscular fat (Martínez-Montes et al., 2016)
	rs81390106	6	88,751,010	*TMEM39B*	Intramuscular Fat (Cesar et al., 2018)
	rs81390106	6	88,751,010	*TXLNA*	Meat quality (Ropka-Molik et al., 2018)
	rs81390106	6	88,751,010	*HDAC1*	Altitude (Ban et al., 2015)
	rs81390106	6	88,751,010	*MARCKSL1*	Feed intake (Lindholm-Perry et al., 2016)
	rs81317489	6	89,640,457	*ZSCAN20*	Scrotal circumference (Sweett et al., 2018)
	rs81317489	6	89,640,457	*CSMD2*	Meat pH trait (Dong et al., 2014); Body weight (Yoshida et al., 2017)
	rs80894853	9	78,663,586	*NXPH1*	DMI (Olivieri et al., 2016)
	rs81389936	6	88,264,983	*COL16A1*	Carcass and meat quality traits (Choi et al., 2012)
	rs80790807	4	106,750,789	*PTPN22*	Immune response (Lamsyah et al., 2009)
	rs80790807	4	106,750,789	*BCL2L15*	Mastitis (Chen et al., 2015)
	rs80911350	14	11,345,116	*SCARA3*	Meat quality traits (Tizioto et al., 2015)
	rs80911350	14	11,345,116	*CLU*	Fertility (Kumar et al., 2015), intramuscular fat (de Jager et al., 2013)
	rs343528814	13	36,608,977	*CACNA2D3*	Reproductive traits (Smith et al., 2019); body width in gilts and sows (Rothschild, 2010), body weight traits (Borowska et al., 2017), altitude (Zhang et al., 2014)
	rs81478390	13	53,707,241	*RYBP*	Body conformation traits - body weight, body length, body height, and chest circumference (Zhou et al., 2016)
	rs81330369	9	7,449,894	*FCHSD2*	Milk production traits (Kemper et al., 2015)
	rs80975991	7	33,481,446	*ZFAND3*	Growth and carcass quality traits (Li and Kim, 2015)
	rs80855522	4	11,0552,282	*GNAI3*	Heat tolerance (Berihulay et al., 2019)
	rs80988392	1	213,780,848	*PTPRD*	Meat quality (Raschetti et al., 2013)

due to Wahlund effect. As expected, we found that the village pig populations of South Africa had high inbreeding values compared with other populations. The negative F_{IS} values for commercial and indigenous populations are reflective of their intensive production environment as individuals are outbred to avoid mating to close relatives.

The low levels of effective population size (N_e) in the recent 12–22 generations for both commercial and indigenous populations are of concern (**Supplementary Table S1**). More so in the indigenous breeds since low levels of genetic diversity are likely to diminish overtime and increase the risk of extinction. The effective population of the Kolbroek of 34 at 12 generations ago is even lower than the minimum threshold N_e of 50 set by the FAO (2000). Franklin (1980) recommended a N_e of at least more than 500 while Willi et al. (2006) suggested N_e of more than 1,000 to maintain the evolutionary potential of any population. The genetic diversity of these populations will likely continue to be negatively impacted by the small number of founders and them being farmed in fragmented populations. Small effective population size of the Kolbroek might be due to pigs being raised in a research facility with limited boars and sows. Large White, Duroc and South African Landrace are commercial pigs that have undergone strong selection for meat and carcass traits thus resulting in small effective population sizes. Long-term sustainability of the populations might be compromised due to the small population size as it increases the effects of genetic drift and reduction in fitness traits (Frankham et al., 1998).

The high F_{IS} values observed within populations across breeds are similar to previous studies (SanCristobal et al., 2006; Swart et al., 2010; Gama et al., 2013; Edea et al., 2014). An overall AMOVA F_{IS} value of 93.95% was comparable to Halimani et al. (2012) value of 92.90% in indigenous pigs of Southern Africa. Diversity amongst South African populations that ranged from F_{CT} = 0.92 (village pigs) to F_{CT} = 5.42 (Commercial populations) might be due to gene flow between different populations within a sub-populations. Moderate diversity within population (i.e., F_{IS} ranging from 19.92 in the category consisting of South African Wild Boar and worldwide Wild Boar to F_{IS} = 35.52 in the categories consisting on South African villages and Worldwide villages) relative to elevated F_{CT} in the same categories implies a higher genetic variation distributed among groups from different geographic locations. This genetic variation observed amongst groups of the South African and Burgos-Paz et al. (2013) pig populations (i.e., F_{CT} = 62.35–73.58) is higher than the variation reported amongst Angora goats from South Africa, France and Argentina using 50K SNP BeadChip (Visser et al., 2016), which could be explained by limited exchange of breeding animals across geographic boundaries in the studied pig populations. The amongst population within groups diversity values ranging from F_{SC} = 0.46 for South African villages to F_{SC} = 18.17 for South African commercial demonstrates evidence of population sub-structure and genetic differentiation between the well-defined commercial and indigenous breeds relative to non-descript village populations that are characterized by weak population boundaries.

The PCA demonstrates the impact of domestication and geographic history on the clustering of populations. European populations as represented by Wild Boar, South African Landrace, and Large White, clustered together as expected (**Figure 3**). Considering the history that the Wild Boar is an ancestor to the domestic pigs of today, some gene flow may have remained from the Wild Boar in the domestic pigs (Giuffra et al., 2000). The clustering of the Wild Boars reflects a European ancestry of those populations within that cluster. The slight difference between the Wild Boar and domestic populations might have been due to geographic isolation and artificial selection. Geographic structures were evident amongst most of the pig populations that were aligned to production systems and their founder effects. The clustering of the Windsnyer and the village populations could be due to gene flow between indigenous breeds and village populations. Limpopo populations had a closer proximity to Large White and South African Landrace, and farmers in this region are more likely to buy pigs from commercial herds. The Large White and South African Landrace are also closer together as these are both European breeds. It was interesting that generally the village populations were closer to the Windsnyer and Kolbroek as these are both indigenous breeds in South Africa. Although not much is known about our indigenous breeds, different theories suggest that the Kolbroek might have far Eastern alleles while the Windsnyer is known to be dominant in other parts of Southern Africa like Mozambique, Zambia and Zimbabwe (Holness, 1973, 1991). The village populations and other Large Whites and Landraces from the global data set clustered together with the South African village, commercial and indigenous pigs demonstrating genetic similarities that could be aligned to founder effects and similarities in production systems.

The clustering of Duroc away from other commercial populations (Large White and South African Landrace) was expected. The Duroc breed was created in the United States with pigs of several ancestries, including African pigs (Porter, 1993). Studies conducted by Kotze and Visser (1996) and Swart et al. (2010) using the microsatellite markers on the Large White, South African Landrace and Duroc also reported similar results. The Large White and South African Landrace were more genetically similar when compared to the Duroc. The inclusion of global populations did not alter this clustering (**Figure 4**).

The distance of Vietnamese Potbelly population from the rest of the domestic pigs is clear evidence of independent domestication that took place between the European and Asian subspecies of the wild boar (Giuffra et al., 2000). The PCA including pigs genotyped from all over the world clearly shows the geographical effect of the populations as the Vietnamese Potbelly clustered in close proximity to the Chinese population.

ADMIXTURE $K = 2$ presented the first level of ancestry of the Suidae family representing *Phacochoerus africanus* (Warthog) and *Potamochoerus larvatus* (Bush pig) versus *Sus scrofa* (domesticated pigs including the Wild Boar) species (**Figure 5**). The presence of the Wild Boar genomic signature in the domestic pigs from $K = 2$ to $K = 7$ is not surprising (**Figure 5**). It is well documented that the domestic pigs diverged from each other and originated from the ancestral wild boars around 8,000–10,000 years ago (Giuffra et al., 2000; Laval et al., 2000; Larson et al., 2005). The Asian and European ancestral wild boars also originated from different subspecies thus the Vietnamese Potbelly diverged early ($K = 2$) from the rest of the domestic pig population. The results for the village populations showed high levels of admixture and weak between population sub-structuring. As opposed to pigs from the commercial sector that practices the intensive production systems, pigs in the villages are farmed under semi-intensive of free-range production systems, which might explain the admixture observed in this study. There is considerable indiscriminate crossbreeding that is taking place in village populations (Rege and Gibson, 2003). European and Asian pigs were used to improve the South African pig breeds but the actual contribution is unknown. Although phenotypically distinct from each other, the Bush pigs and warthogs clustered together which is suggestive of either common founder effect or selection pressures in the natural environments.

According to Wright (1978), F_{ST} estimation with values of less than 0.05 represents low differentiation while values between 0.05 and 0.15 represent a moderate genetic differentiation and those between 0.15 and 0.25 and beyond reflect highly differentiated populations. The low levels of genetic differentiation of the village populations from this study (**Table 3**) is consistent to pairwise F_{ST} values of Halimani et al. (2012) of village populations from Zimbabwe and South Africa. Most pig farmers from the villages practice free ranging or semi-controlled farming where there is continuous gene flow between populations within villages thereby explaining the low levels of population sub-structuring observed. Moderate F_{ST} values implies closer relationship between the South African Landrace and Large White and agrees with their breeding history, whereby the Landrace was developed from crossing the Large White from England and a Denmark indigenous. Greater genetic differentiation between the Warthog and the other pig populations (F_{ST} = 0.36–0.53) might be attributed to the (i) pressures of natural selection (ii) the separate histories of domestic and wild populations and (iii) the unique population dynamics of Warthogs that are known to live in clans of adult females, males and their offspring while maintaining minimal contacts with other clans (Cumming, 1975; Somers et al., 1994). In South Africa, Warthog populations are restricted to nature reserves thus creating a physical barrier and huge genetic differentiation between them and other pig populations. This will be in contrast to the greater interaction between village, commercial and indigenous populations. Low F_{ST} values between the villages in South African and village populations from South America (**Supplementary Table S3**) from Burgos-Paz et al. (2013) study, might be an indication that either common founder populations or similarities in production systems leading to common selection pressures. Ramírez et al. (2009) demonstrated that the African and South American pigs were derived from Europe and Far Eastern pigs. The very high genetic differentiation between the Vietnamese Potbelly and Bush pig agrees with the PCA and Admixture clustering.

Per marker pairwise F_{ST} were estimated between pairs highly differentiated populations which were from villages, commercial, indigenous, Asian and wild populations (**Table 3**).

From the pairwise F_{ST}, Warthog was found to be genetically different from the rest of the populations. The per marker pairwise F_{ST} analysis used a threshold of 0.8 and above to plot Manhattan graphs of the Warthog against the rest of the populations. From the SNPs showing a threshold of $F_{ST} \geq 0.8$, we looked at candidate genes and QTLs that can be associated with those SNPs to infer on traits that might have genetically differentiated the Warthog from Alfred Nzo, Duroc, Kolbroek, Large White, South African Landrace, and Windsnyer populations (**Supplementary Figure S2**).

Majority of the SNPs that were above the threshold between the Warthog and the rest of the populations were from chromosomes 1, 4, 5, 12, 13, and 15 (**Table 4**). Chromosomes 2 (Warthog vs. Alfred Nzo), 3 (Warthog vs. Kolbroek), 6 (Warthog vs. South African Landrace) and 14 (Warthog vs. Large White) seemed to be less common. Chromosome 1 with a total number of 12 SNPs was associated with reproduction and growth traits while the indigenous populations of Kolbroek and Windsnyer were differentiated on chromosome 4 that was also linked to reproduction and growth traits.

Warthog vs. Alfred Nzo had three SNPS ($F_{ST} \geq 0.8$) that are associated with reproduction (*RPL18*, *IL17B*) and growth (*IL17B*, *ARHGAP23*) characteristics (**Table 4**). It is known that good nutrition is vital to be able to maximize growth performance. Genes *IL17B* and *ARHGAP23* are linked to inflammatory response (Liu, 2015; Bie et al., 2017) and the gastrointestinal tract where they play a role in the digestion and absorption of the nutrients. Inflammatory responses lead to reduction of feed intake, which in turn affects the growth of the animal (Liu, 2015). Selection on genes associated with inflammation in the populations of Warthog vs. Alfred Nzo might be an effect of the different diets these populations scavenge on. Medzhitov (2008) noted the inflammation response to be a protective mechanism from the stress and harmful environment.

Growth linked genes *ADGRB3*, and *ACY1* were dominant in differentiating Warthog vs. Duroc populations with an overall total of 10 SNPs. Emrani et al. (2017) associated *ADGRB3* to body weight traits in the broiler chickens. The association of *ADGRB3* gene to Duroc rather than Large White or South African Landrace breeds might be linked to the higher percentage of intramuscular fat in Duroc compared to the other two commercial breeds (De Vries et al., 2000). Mature males of Warthog can also reach up to 100 kg and possesses good meat and carcass qualities (Hoffman and Sales, 2007).

A total number of 20 significant SNPs ($F_{ST} \geq 0.8$) were linked to the Warthog vs. Kolbroek populations. Growth traits were associated with five of the SNPs between Warthog vs. Kolbroek. Indigenous Kolbroek are reported to be smaller in size when compared to commercial breeds such as Large White (Chimonyo et al., 2005). Kutwana et al. (2015) reported no significant difference ($P > 0.05$) between the Kolbroek and Large White populations that had higher fat percentages when compared to the other commercial breeds (Nicholas, 1999).

Chromosome 13 was also highly notable with significant SNPs differentiating Warthog vs. Kolbroek and Warthog vs. Windsnyer. Only two SNPs appeared for Warthog vs. South African Landrace and were on chromosome 6. The Warthog vs. Windsnyer had a total of fourteen SNPs differentiating them. The identification of BRPF1 gene in the Warthog vs. Windsnyer populations is an important observation as this gene is associated with the intramuscular fat (IMF). When it comes to the value and taste of the pork meat, intramuscular fat is an important characteristic because meat that is high in IMF tends to be juicy and tender (Eikelenboom et al., 1996; de Koning et al., 1999). The gene *ATPB2* associated with six significant SNPs is linked to heat stress and reproductive performance (Dash et al., 2016). Heat stress might result in poor reproduction for both sows and boars. Pigs cannot sweat and this makes them sensitive to high environmental temperatures making and of concern particularly to commercial pig farmers (Ross et al., 2015).

Genes linked to immune response and mastitis were observed in Indigenous vs. Duroc comparisons. *PTPN22* gene on chromosome 4 has a regulatory effect on T- and B-cell activation in immune response (Lamsyah et al., 2009). *PTPN22* plays a role in susceptibility to tuberculosis. Pigs are generally natural hosts of mycobacterial infections (de Lisle, 1994). Porcine TB has been reported in South Africa where infections are commonly via infected cattle fecal matter fed to piglets as well as interactions with wild pigs (Muwonge et al., 2012). *NXPH1* gene is associated with DMI (dry matter intake) in cattle (Olivieri et al., 2016). Both *PTPN* and *NXPH1* genes were fixed in the Duroc implying natural selection of the Duroc when compared to both indigenous and Wild Boars. Breeds in the commercial sector are mainly selected for growth, carcass and meat quality traits. The indigenous and village population on the other hand has not been systematically selected for such traits.

The *NPY5R* located on chromosome 8, was associated with feed efficiency and fat deposition. This gene was also reported in Jinhua and Rongchang pigs that belong to Chinese breeds (Chen et al., 2018). Fat deposition genes observed in Indigenous vs. Vietnamese, Villages vs. Kolbroek and South African Landrace with Large White vs. Indigenous are evidence in agreement with suggestions that Kolbroek and other indigenous pigs tend to carry their weight in their bellies and backs (Hoffman et al., 2005). Hoffman et al. (2005) also reported breed type and diet to have an influence on the composition of the meat. This study therefore presented a diverse genomic architecture of South African pigs with differentiating selection pressures for meat and carcass quality traits in the different pigs raised in diverse production systems.

CONCLUSION

Overall, the study demonstrated the utility of the Porcine SNP60K BeadChip in elucidating genetic diversity and population genomic structure of South African pig populations relative to other global populations. Village pigs demonstrated distinctiveness from other domestic and commercial populations

within South Africa and when compared to global populations. The study provided baseline knowledge with regards to the genetic diversity of the domestic and wild pig populations of South Africa, which is a prerequisite for population/breed characterization, utilization and conservation. A more in-depth analysis of patterns of genetic variations is required to get more insight into factors shaping genetic diversity of these populations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in Dryad doi: 10.5061/dryad.b0t10b0.

ETHICS STATEMENT

Ear tissue samples were collected from pigs using the Tissue Sampling Applicator Gun while pliers were used to collect the hair samples according to standard procedures and ethical approval from ARC-Irene Animal Ethics committee (APIEC16/028).

AUTHOR CONTRIBUTIONS

NH collected samples, analyzed the data, and wrote the draft manuscript. FM, PS, and ED designed the experiment and sourced funding. KH analyzed the genomic data for the experiment. FM, PS, and ED coordinated the conduct of the study and writing of manuscript and revisions. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

We would like to thank the Agricultural Research Council-Biotechnology Platform (ARC-BTP) for funding the genotyping of samples. We express our gratitude to all the pig farmers, the Department of Agriculture (Eastern Cape and Limpopo) and various stakeholders who allowed us to use their animals in this study. NH holds fellowships from the National Research Foundation and ARC-Professional Development Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00344/full#supplementary-material

FIGURE S1 | Cross validation plot for inferring the number of *K* populations in the analysis of population structure.

FIGURE S2 | Genome-wide Manhattan plot of F_{ST} among the pig populations of **(a)** Alfred Nzo and Warthog, **(b)** Duroc and Warthog, **(c)** Kolbroek and Warthog, **(d)** Large White and Warthog, **(e)** South African landrace and Warthog, **(f)** Windsnyer and Warthog, **(g)** Indigenous and Duroc, **(h)** Villages and Duroc, **(i)** Villages and Kolbroek, **(j)** South African Landrace with Large White and Indigenous, **(k)** Indigenous and Vietnamese, **(l)** Villages and Wild Boar, **(m)** Villages and Vietnamese, **(n)** Wild Boar and Duroc. The solid lines indicate the $F_{ST} \geq 0.8$ thresholds.

TABLE S1 | Average effective population size estimates across generations for the different populations analyzed.

TABLE S2 | Partitioning of genetic variance for the different populations analyzed.

TABLE S3 | F_{ST} between SA populations and genotypes from Burgos-Paz et al. (2013).

REFERENCES

Ai, H., Huang, L., and Ren, J. (2013). Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. *PLoS One* 8:e56001. doi: 10.1371/journal.pone.0056001

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246. doi: 10.1186/1471-2105-12-246

Amills, M., Clop, A., Ramírez, O., and Pérez-Enciso, M. (2010). "Origin and diversity of pig breeds," in *Encyclopedia of Life Sciences*, ed. H. Kehrer-Sawatzki (Chichester: John Wiley & Sons), 1–7. doi: 10.1002/9780470015902.a002288

Amills, M., Ramírez, O., Galman-Omitogun, O., and Clop, A. (2012). Domestic pigs in Africa. *Afr. Archaeol. Rev.* 30, 73–82. doi: 10.1007/s10437-012-9111-2

Ban, D. M., Zhang, B., Wang, Z. X., Zhang, H., and Wu, C. X. (2015). Differential gene expression of epigenetic modifying enzymes between Tibet pig and Yorkshire in high and low altitudes. *Genet. Mol. Res.* 14, 3274–3280. doi: 10.4238/2015.April.13.6

Berihulay, H., Abied, A., He, X., Jang, L., and Ma, Y. (2019). Adaptation mechanisms of small ruminants to environmental heat stress. *Animals* 9:75. doi: 10.3390/ani9030075

Bie, Q., Jin, C., Zhang, B., and Dong, H. (2017). IL-17B: a new area of study in the IL-17 family. *Mol. Immunol.* 90, 50–56. doi: 10.1016/j.molimm.2017.07.004

Blench, R. M. (2000). "A history of pigs in Africa," in *Origins and Development of African Livestock: Archaeology, Genetics, Linguistics and Ethnography*, eds R. M. Blench, and K. Mac Donald (Abingdon: Routledge Books).

Blench, R. M., and MacDonald, K. C. (eds) (2000). "The origins and development of the African livestock,"in *Archaelogy, Genetics, Linguistics and Ethnography*. London: UCL Press

Borowska, A., Reyer, H., Wimmers, K., Varley, P. and Szwackowski, T. (2017). Detection of pig genome regions determining production traits using an information theory approach. *Livest. Sci.* 205, 31–35. doi: 10.1016/j.livsci.2017.09.012

Browett, S., McHugo, G., Richardson, I. W., Magee, D. A., Park, S. D. E., Fahey, A. G. et al. (2018). Genomic characterisation of the indigenous Irish Kerry cattle breed. *Front. Genet.* 9:51. doi: 10.3389/fgene.2018.00051

Burgos-Paz, W., Souza, C. A., Megens, H. J., Ramayo-Caldas, Y., Melo, M., Lemús-Flores, C., et al. (2013). Porcine colonization of the Americas: a 60k SNP story. *Hereditary* 110, 321–330. doi: 10.1038/hdy.2012.109

Cesar, A. S., Regitano, L. C., Reecy, J. M., Poleti, M. D., Oliveira, P. S. N., De Oliveira, G. B., et al. (2018). Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. *BMC Genomics* 19:499. doi: 10.1186/s12864-018-4871-y

Chen, M., Wang, J., Wang, Y., Wu, Y., Fu, J., and Liu, J. F. (2018). Genome-wide detection of selection signatures in Chinese indigenous Laiwu pigs revealed candidate genes regulating fat deposition in muscle. *BMC Genet.* 19:31. doi: 10.1186/s12863-018-0622-y

Chen, X., Cheng, Z., Zhang, S., Werling, D., and Wathes, D. C. (2015). Combining genome wide association studies and differential gene expression data analyses identifies candidate genes affecting mastitis caused by two different pathogens in the dairy cow. *Open J. Anim. Sci.* 5, 358–393. doi: 10.4236/ojas.2015.54040

Chimonyo, M., Bhebhe, E., Dzama, K., Halimani, T. E., and Kanengoni, A. (2005). Improving smallholder pig production for food security and livelihoods of the poor in Southern Africa. *Proc. Afr. Crop Sci. Conf.* 7, 569–573.

Choi, I., Bates, R. O., Raney, N. E., Steibel, J. P., and Ernst, C. W. (2012). Evaluation of QTL for carcass merit and meat quality traits in a US commercial Duroc population. *Meat Sci.* 92, 132–138. doi: 10.1016/j.meatsci.2012.04.023

Corbin, L. J., Liu, A. Y., Bishop, S. C., and Woolliams, J. A. (2012). Estimation of historical effective population size using linkage disequilibrium with marker data. *J. Anim. Breed. Genet.* 129, 257–270. doi: 10.1111/j.1439-0388.2012.01003.x

Cornelis, M. C., and Hu, F. B. (2013). Systems epidemiology: a new direction in nutrition and metabolic disease research. *Curr. Nutr. Rep.* 2, 225–235. doi: 10.1007/s13668-013-0052-4

Cumming, D. H. M. (1975). *A Field of Study of the Ecology and Behaviour of Warthog*. Harare: National Museums and Monuments of Rhodesia,

D'Huart, J., and Grubb, P. (2003). Distribution of the common warthog (*Phacochoerus africanus*) and the desert warthog (*Phacochoerus aethiopicus*) in the Horn of Africa. *Afr. J. Ecol.* 39, 156–169. doi: 10.1046/j.0141-6707.2000.00298.x

Dash, S., Chakravarty, A. K., Singh, A., Upadhyay, A., Singh, M. and Yousef, S. (2016). Effect of heat stress on reproductive performances of dairy cattle and buffaloes: a review. *Vet. World* 9, 235–244. doi: 10.14202/vetworld.2016.235-244

de Jager, N., Hudson, N. J., Reverter, A., Barnard, R., Café, L. M., Greenwood, P. L., et al. (2013). Gene expression phenotypes for lipid metabolism and intramuscular fat in skeletal muscle of cattle. *J. Anim. Sci.* 91, 1112–1128. doi: 10.2527/jas.2012-5409

de Koning, D. J., Janss, L. L., Rattubk, P., van Oers, P. A., de Vries, B. J., Groenen, M. A., et al. (1999). Detection of quantitative traits loci for backfat thickness and intramuscular fat content in pig. *Genetics* 152, 1679–1690

de Lisle, G. W. (1994). Mycobacterial infections in pigs. *Surveillance* 21, 23–25.

De Vries, A. G., Faucitano, L., Sosnicki, A., and Plastow, G. H. (2000). The use of gene technology for optimal development of pork meat quality. *Food Chem.* 69, 397–405. doi: 10.1016/S0308-8146(00)00049-2

Dong, Q., Liu, H., Li, X., Wei, W., Zhao, S., and Cao, J. (2014). A genome-wide association study of five meat quality traits in Yorkshire pigs. *Front. Agr. Sci. Eng.* 1, 137–143. doi: 10.15302/J-FASE-2014014

Du, Y., Li, M., Chen, J., Duan, Y., Wang, X., Qiu, Y., et al. (2016). Promoter targeted bisulfite sequencing reveals DNA methylation profiles associated with low sperm motility in asthenozoospermia. *Hum. Reprod.* 31, 24–33. doi: 10.1093/humrep/dev283

Edea, Z., Bhuiyan, M. S. A., Dessie, T., Rothschild, M. F., Dadi, H., and Kim, K. S. (2014). Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. *Animal* 9, 218–226. doi: 10.1017/S1751731114002560

Edea, Z., Hong, J. K., Jung, J. H., Kim, D. W., Kim, E. S., Shin, S. S., et al. (2017). Detecting selection signatures between Duroc and Duroc synthetic pig populations using high-density SNP chip. *Anim. Genet.* 48, 473–477. doi: 10.1111/age.12559

Eikelenboom, G., Hoving-Bolink, A. H., and Van Der Wal, P. G. (1996). The eating quality of pork: 2. The influence of intramuscular fat. *Fleischwirtschaft* 76, 517–518.

Emrani, H., Torshizi, R.V., Masoudi, A.A., and Ehsani, A. (2017). Identification of new loci for body weight traits in F2 chicken population using genome-wide association study. *Livest. Sci.* 206, 125–131. doi: 10.1016/j.livsci.2017.10.016

Eusebi, P. G., Cortés, O., Carleos, C., Dunner, S., and Cañon, J. (2018). Detection of selection signatures for agonistic behaviour in cattle. *J. Anim. Breed. Genet.* 135, 170–177. doi: 10.1111/jbg.12325

Excoffier, L., Laval, G., and Schneider, S. (2005). ARLEQUIN ver. 3.0: an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* 1, 47–50.

Fang, Z. H., and Pausch, H. (2019). Multi-trait meta-analyses reveal 25 quantitative trait loci for economically important traits in Brown Swiss cattle. *bioRxiv* [Preprint]. doi: 10.1101/517276

FAO (2000). *Secondary Guidelines of Farm Animal Genetic Resources Management Plans. Management of Small Populations at Risk*. Rome: Food and Agriculture Organization of the United Nations.

FAO (2009). *Status and Trends Report on Animal Genetic Resources–2008*. Rome: Food and Agriculture Organization of the United Nations.

Ferreira, S., Gaylard, A., Greaver, C., Hayes, J., Cowell, C., and Ellis, G. (2013). *Summary Report: Animal abundances in Parks 2012/2013*. Skukuza: SANParks.

Fischer, D., Laiho, A., Gyenesei, A., and Sironen, A. (2015). Identification of reproduction-related gene polymorphisms using whole transcriptome sequencing in the Large White pig population. *G3* 5, 1351–1360. doi: 10.1534/g3.115.018382

Fontanesi, L., Schiavo, G., Galimberti, G., Calò, D. G., and Russo, V. (2014). A genomewide association study for average daily gain in Italian Large White pigs. *J. Anim. Sci.* 92, 1385–1394. doi: 10.2527/jas2013-7059

Frankenberg, S. R., De Barros, F. R. O., Rossant, J., and Renfree, M. B. (2016). The mammalian blastocyst. *J. Dev. Biol.* 5, 210–232. doi: 10.1002/wdev.220

Frankham, R., Ballou, J. D., and Briscoe, D. A. (1998). *An Introduction to Conservation Genetics*. Cambridge: Cambridge University Press.

Franklin, I. R. (1980). "Evolutionary change in small populations," in *Conservation Biology: An Evolutionary-Ecological Perspective*, eds M. Soulé, and B. Wilcox (Sunderland, MA: Sinauer Associates).

Gaddis, K. P., Null, D. J., and Cole, J. B. (2016). Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. *J. Dairy Sci.* 99, 6420–6435. doi: 10.3168/jds.2015-10444

Gama, L. T., Martínez, A. M., Carolino, I., Landi, V., Delgado, J. V., and Vicente, A. A. (2013). Genetic structure, relationships and admixture with wild relatives in native pig breeds from Iberia and its island. *Genet. Sel. Evol.* 45:18. doi: 10.1186/1297-9686-45-18

Giuffra, E., Kijas, J. M., Amarger, V., Carlborg, O., and Andersson, L. (2000). The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* 154, 1785–1791

Gmelin, J. F. (1788). *Systema Naturae per Regna Tria Naturae Secundum Classes, Ordines, Genera, Species Cum Caracteribus, Differentiis, Synonmis, Locis*, Vol. 1. Leipzig: CRC Press.

Gondret, F., Vincent, A., Houée-Bigot, M., Siegel, A., Lagarrigue, S., Causeur, D., et al. (2017). A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. *BMC Genomics* 18:244. doi: 10.1186/s12864-017-3639-0

Gray, C. A., Abbey, C. A., Beremand, P. D., Choi, Y., Farmer, J. L., Adelson, D. L., et al. (2006). Identification of endometrial genes regulated by early pregnancy, progesterone, and interferon tau in the ovine uterus. *Biol. Reprod.* 74, 383–394. doi: 10.1095/biolreprod.105.046656

Grossi, D. A., Abo-Ismail, M. K., Koeck, A., Miller, S. P., Stothard, P., Plastow, G., et al. (2014). "Genome-wide association analyses for mastitis in Canadians Holsteins," in *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*, Wageningen.

Gu, T., Zhu, M., Schroyen, M., Qu, L., Nettleton, D., Kuhar, D., et al. (2014). Endometrial gene expression profiling in pregnant Meishan and Yorkshire pigs on day 12 of gestation. *BMC Genomics* 15:156. doi: 10.1186/1471-2164-15-156

Gutiérrez-Gil, B., Wiener, P., and Williams, J. L. (2007). Genetic effects on coat colour in cattle: dilution of eumelanin and phaeomelanin pigments in an F2-backcross Charolais × Holstein population. *BMC Genomics* 8:56. doi: 10.1186/1471-2156-8-56

Halimani, T. E., Muchadeyi, F. C., Chimonyo, M., and Dzama, K. (2012). Some insights into the phenotypic and genetic diversity of indigenous pigs in Southern Africa. *S. Afr. J. Anim. Sci.* 42, 507–510. doi: 10.4314/sajas.v42i5.1

Hamill, R. M., McBryan, J., McGee, C., Mullen, A. M., Sweeney, T., Talbot, A., et al. (2012). Functional analysis of muscle gene expression profiles associated with tenderness and intramuscular fat content in pork. *Meat Sci.* 92, 440–450. doi: 10.1016/j.meatsci.2012.05.007

Hatzirodos, N., Hummitzsch, K., Irving-Rodgers, H. F., Harland, M. L., Morris, S. E., and Rodgers, R. J. (2014a). Transcriptome profiling of granulosa cells from bovine ovarian follicles during atresia. *BMC Genomics* 15:40. doi: 10.1186/1471-2164-15-40

Hatzirodos, N., Irving-Rodgers, H. F., Hummitsch, K., Harland, M. L., Morris, S. E., and Rodgers, R. J. (2014b). Transcriptome profiling of granulosa cells of bovine ovarian follicles during growth from small to large antral sizes. *BMC Genomics* 15:24. doi: 10.1186/1471-2164-15-24

Hunt, S. C., Hasstedt, S. J., Xin, Y., Dalley, B. K., Milash, B. A., Yakobson, E., et al. (2011). Polymorphisms in the *NPY2R* Gene Show Significant Associations with BMI that are Additive to *FTO*, *MC4R*, and *NPFFR2* Gene Effects. *Obesity* 19, 2241–2247. doi: 10.1038/oby.2011.239

Mdladla, K., Dzomba, E. F., Huson, H., and Muchadeyi, F. C. (2016). Population genomic structure and linkage disequilibrium analysis of South African goat breeds using genome-wide SNP data. *Anim. Genet.* 47, 471–482. doi: 10.1111/age.12442

Hoffman, L. C., and Sales, J. (2007). Physical and chemical quality characteristics of warthog (*Phacochoerus africanus*) meat. *Livest. Res. Rural Dev.* 19:153 doi: 10.1016/j.meatsci.2018.07.001

Hoffman, L. C., Styger, W. F., Brand, T. S., and Muller, M. (2005). The growth, carcass yield, physical and chemical characteristic of two South African indigenous pig breeds. *S. Afr. J. Anim. Sci.* 6, 25–35

Holness, D. H. (1973). The role of indigenous pigs as source of protein in Africa: a review. *Rhode. Agric. J*, 73, 59–63.

Holness, D. H. (1991). *The Tropical Agriculturalist*. London: Macmillan Education Ltd.

Ibeagha-Awemu, E. M., Peters, S. O., Akwanji, K. A., Imumorin, I. G., and Zhao, X. (2016). High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. *Sci. Rep.* 6:31109. doi: 10.1038/srep31109

Jones, G. F. (1998). "Genetic aspects of domestication, common breeds and their origin," in *The Genetics of the Pig*, eds M. F. Rothschild, and A. Ruvinsky (Wallingford: CABI Publishing).

Jori, F., and Bastos, D. S. (2009). Role of wild suids in the epidemiology of African swine fever. *Ecohealth* 6, 296–310. doi: 10.1007/s10393-009-0248-7

Júnior, G. F., Costa, R. B., De Camargo, G. M., Carvalheiro, R., Rosa, G. J., Baldi, F., et al. (2016). Genome scan for postmortem carcass traits in Nellore cattle. *J. Anim. Sci.* 94, 4087–4095. doi: 10.2527/jas2016-0632

Kang, K., Seo, D. W., Lee, J. B., Jiung, E., Park, H., Cho, I., et al. (2013). Identification of SNPs affecting porcine carcass weight with the 60K SNP chip. *J. Anim. Sci. Technol.* 55, 231–235. doi: 10.5187/JAST.2013.55.4.231

Kem, E. H. (ed.) (1993). *Pig Production in South Africa*. Irene: Agricultural Research Council.

Kemper, K. E., Reich, C. M., Bowman, P. J., van der Jagt, C. J., Chamberlain, A. J., Mason, B. A., et al. (2015). Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genet. Sel. Evol.* 47:29. doi: 10.1186/s12711-014-0074-4

Khanyile, K. S., Dzomba, E. F., and Muchadeyi, F. C. (2015). Population genetic structure, linkage disequilibrium and effective population size of conserved and extensively raised chicken populations of Southern Africa. *Front. Genet.* 6:13. doi: 10.3389/fgene.2015.00013

Kimura, A., Namekata, K., Guo, X., Harada, C., and Harada, T. (2016). Neuroprotection, growth factors and BDNF-TrkB signalling in retinal degeneration. *Int. J. Mol. Sci.* 17:1584.

Kotze, A., and Visser, D. P. (1996). "Status of genetic variation in purebred pig breeds in South Africa," in *Proceedings of the 7th All Africa Conference on Animal Agriculture*, Pretoria.

Krige, J. E. (1950). *The Social Systems of the Zulus*, 2nd Edn. Pietermaritzburg: Shuter and Shooter.

Kumar, S., Deb, R., Singh, U., Ganguly, I., Mandal, D. K., Tyagi, S., et al. (2015). Bovine circadian locomotor output cycles kaput (CLOCK) and clusterin (CLU) mRNA quantitation in ejaculated crossbred bull spermatozoa. *Reprod. Domest. Anim* 50, 505–509. doi: 10.1111/rda.12522

Kutwana, H. W., Gxasheka, M., and Tyasi, T. L. (2015). Body weight and morphological traits of Large White and Kolbroek pig breeds. *Int. J. Adv. Res.* 3, 105–109.

Lamsyah, H., Rueda, B., Baassi, L., Elaouad, R., Bottini, N., Sadki, K., et al. (2009). Association of PTPN22 gene functional variants with development of pulmonary tuberculosis in Moroccan population. *Tissue Antigens* 74, 228–232. doi: 10.1111/j.1399-0039.2009.01304.x

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* 307, 1618–1621. doi: 10.1126/science.1106927

Laval, G., Iannuccelli, N., Legault, C., Milan, D., Groenen, M. A., Giuffra, E., et al. (2000). Genetic diversity of eleven European pig breeds. *Genet. Sel. Evol.* 32, 187–203. doi: 10.1186/1297-9686-32-2-187

Le, T. H., Christensen, O. F., Nielsen, B., and Sahana, G. (2017). Genome-wide association study for conformation traits in three Danish pig breeds. *Genet. Sel. Evol.* 49:12. doi: 10.1186/s12711-017-0289-2

Lee, W., Taye, M., Kwon, T., Yoon, J., Jang, D., Suzuki, S., et al. (2017). Identifying candidate positive selection genes in Korean imported pig breeds. *Genes Genomics* 39, 557–565. doi: 10.1007/s13258-017-0529-4

Li, X., Kim, S. W., Choi, J. S., Lee, Y. M., Lee, C. K., Choi, B. H., et al. (2010). Investigation of porcine FABP3 and LEPR gene polymorphisms and mRNA expression for variation in intramuscular fat content. *Mol. Biol. Rep.* 37, 3931–3939. doi: 10.1007/s11033-010-0050-1

Li, X., Kim, S.-W., Do, K.-T., Ha, Y.-K., Lee, Y.-M., Yoon, S.-H., et al. (2011). Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies. *Mol. Biol. Rep.* 38, 3805–3820. doi: 10.1007/s11033-010-0496-1

Li, Y., and Kim, J. J. (2015). Multiple linkage disequilibrium mapping methods to validate additive quantitative trait loci in Korean native cattle (Hanwoo). *Asian Australas J. Anim. Sci.* 28, 926–935. doi: 10.5713/ajas.15.0077

Lim, D., Lee, S., Kim, N., Cho, Y., Chai, H., Seong, H., et al. (2013). Gene co-expression analysis to characterize genes related to marbling trait in Hanwoo (Korean) cattle. *Asian-Australas J. Anim. Sci.* 26, 19–29. doi: 10.5713/ajas.2012.12375

Linnaeus, C. (1758). *Systema Naturae per Regna Tria Naturae, secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis. Tomus I. Editio Decima, Reformata*. Stockholm: Laurentii Salvii.

Lindholm-Perry, A. K., Butler, A. R., Kern, R. J., Hill, R., Kuehn, L. A., Wells, J. E., et al. (2016). Differential gene expression in the duodenum, jejunum and ileum among crossbred beef steers with divergent gain and feed intake phenotypes. *Anim. Genet.* 47, 408–427. doi: 10.1111/age.12440

Lionikas, A., Cheng, R., Lim, J. E., Palmer, A. A., and Blizard, D. A. (2010). Fine-mapping of muscle weight QTL in LG/J and SM/J intercrosses. *Physiol. Genomics* 42A, 33–38. doi: 10.1152/physiolgenomics.00100.2010

Lionikas, A., Meharg, C., Derry, J. M. J., Ratkevicius, A., Carroll, A. M., Vandenbergh, D. J., et al. (2012). Resolving candidate genes of mouse skeletal muscle QTL via RNA-Seq and expression network analyses. *BMC Genomics* 13:592. doi: 10.1186/1471-2164-13-592

Liu, Y. (2015). Fatty acids, inflammation and intestinal health in pigs. *J. Anim. Sci. Biotechnol.* 6:41. doi: 10.1186/s40104-015-0040-1

Lönnberg, E. (1908). Remarks on some wart-hog skulls in the British Museum. *Proc. Zool. Soc. London* 78, 936–940. doi: 10.1111/j.1469-7998.1908.00936.x

Lönnberg, E. (1912). Mammals collected by the Swedish zoological expedition to British East Africa 1911. *Kungliga Svenska Vetenskapakademiens Handligar* 48, 1–188.

Luo, W., Cheng, D., Chen, S., Wang, L., Li, Y., and Ma, X. (2012). Genome-wide association analysis of meat quality traits in a porcine Large White × Minzhu intercross population. *Int. J. Biol. Sci.* 8, 580–595. doi: 10.7150/ijbs.3614

Ma, X. J., Zhang, X. L., Wang, L. X., and Liu, Z. (2011). Studies on difference of immune and production indexes between Songliao black pig and large white pig. *China Anim. Husbandry Vet. Med.* 38, 52–55.

Makina, S. O., Muchadeyi, F. C., van Marle-Koöster, E., Macneil, M. D., and Maiwashe, A. (2014). Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. *Front. Genet.* 5:333. doi: 10.3389/fgene.2014.00333

Martínez-Montemayor, M. M., Hill, G. M., Raney, N. E., Rilington, V. D., Tempelman, R. J., Link, J. E., et al. (2008). Gene expression profiling in hepatic tissue of newly weaned pigs fed pharmacological zinc and phytase supplemented diets. *BMC Genomics* 9:412. doi: 10.1186/1471-2164-9-421

Martínez-Montes, A. M., Muiños-Bühl, A., Fernández, A., Folch, J. M., Ibáñez-Escriche, N., and Fernández, A. (2016). Deciphering the regulation of porcine genes influencing growth, fatness and yield-related traits through genetical genomics. *Mamm. Genome* 28, 130–142.

McGraw, K., and List, A. (2017). Chapter five-Erythropoietin receptor signalling and lipid rafts. *Vitam. Horm.* 105, 79–100. doi: 10.1016/bs.vh.2017.02.002

Medzhitov, R. (2008). Origin and physiological roles of inflammation. *Nature* 454, 428–435. doi: 10.1038/nature07201

Messad, F., Louveau, I., Koffi, B., Gilbert, H., and Gondret, F. (2019). Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. *BMC Genomics* 20:659. doi: 10.1186/s12864-019-6010-9

Moioli, B., D'Andrea, S., De Grossi, L., Sezzi, E., de Sanctis, B., Catillo, G., et al. (2016). Genomic scan for identifying candidate genes for paratuberculosis resistance in sheep. *Anim. Prod. Sci.* 56, 1046–1055. doi: 10.1071/ANI14826

Moran, B., Butler, S., Moore, S., MacHugh, D. E., and Creevey, C. J. (2017). Differential gene expression in the endometrium reveals cytoskeletal and

immunological genes in lactating dairy cows genetically divergent for fertility traits. *Reprod. Fertil. Dev.* 29, 274–282. doi: 10.1071/RD15128

Mujibi, F. D., Okoth, E., Cheruiyot, E. K., Onzere, C., Bishop, R. P., Fèvre, E. M., et al. (2018). Genetic diversity, breed composition and admixture of Kenyan domestic pigs. *PLoS One* 13:e0190080. doi: 10.1371/journal.pone.0190080

Muwanika, V. B., Nyakaana, S., Siegismund, H. R., and Arctander, P. (2003). Phylogeography and population structure of the common warthog (*Phacochoerus africanus*) inferred from variation in mitochondrial DNA sequences and microsatellite loci. *J. Hered.* 91, 361–372. doi: 10.1038/sj.hdy.6800341

Muwonge, A., Johansen, T. B., Vigdis, E., Godfroid, J., Olea-Popelka, F., Demelash, B., et al. (2012). Mycobacterium bovis infections in slaughter pigs in Mubende district, Uganda: a public health concern. *BMC Vet. Res.* 8:168. doi: 10.1186/1746-6148-8-168

Naude, R. T., and Visser, D. P. (1994). "ń Generiese kwalitatiewe benadering ten einde die doeltreffende produksie van verbruikersaanneemlike varkvleis te verseker," in *Proceedings of the 6de Nasionale SAVPO-Kongres te Elangeni Hotel*, KwaZulu-Natal.

Newton, J. R., De Santis, C., and Jerry, D. R. (2012). The gene exptression response of the catadromous perciform barramundi Lates calcarifer to an acute heat stress. *J. Fish. Biol.* 81, 81–93. doi: 10.1111/j.1095-8649.2012.03310

Nicholas, G. (1999). *Kolbroek-the Unique Local Breed. Farmer's Weekly*, August.

Novianti, I., Pitchford, W. S., and Bottema, C. D. (2010). Beef cattle muscularity candidate genes. *J. Ilmu Ilmu Peternakan* 20, 1–10.

Olivieri, B. F., Mercadante, M. E., Cyrillo, J. N., Branco, R. H., Bonilha, S. M., Albuquerque, L. G., et al. (2016). Genomic regions associated with feed efficiency indicator traits in an experimental Nellore cattle population. *PLoS One* 11:e0164390. doi: 10.1371/journal.pone.0164390

Pallas, P. S. (1766). *Miscellanea Zoological Quibus Novae Imprimis Atque Obscurae Animalium Species Describuntur et Observationibus Iconibusque Illustrantur*. The Hague: Nabu Press.

Parker Gaddis, K. L., Megonigal, J. H. Jr., Clay, J. S., and Wolfe, C. W. (2018). Genome-wide association study for ketosis in US jerseys using producer-recorded data. *J. Dairy Sci.* 101, 413–424. doi: 10.3168/jds.2017-13383

Piórkowska, K., Żukowski, K., Ropka-Molik, K., Tyra, M., and Gurgul, A. (2018). A comprehensive transcriptome analysis of skeletal muscles in two Polish pig breeds differing in fat and meat quality traits. *Genet. Mol. Biol.* 41, 125–136. doi: 10.1590/1678-4685-GMB-2016-0101

Porter, V. (1993). *Pigs: A Handbook to the Breeds of the World*. Mountfield, HK: Helm Information Ltd.

Puig-Oliveras, A., Revilla, M., Castelló, A., Fernández, A. I., Folch, J. M., and Ballester, M. (2016). Expression-based GWAS identifies variants, gene interactions and key regulators affecting intramuscular fatty acid content and composition in porcine meat. *Sci. Rep.* 6:31803. doi: 10.1038/srep31803

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

Ramírez, O., Ojeda, A., Tomas, A., Gallardo, D., Huang, L. S., Folch, J. M., et al. (2009). Integrating Y-chromosome, mitochondrial, and autosomal data to analyse the origin of pig breeds. *Mol. Biol. Evol.* 26, 2061–2072. doi: 10.1093/molbev/msp118

Ramos, A. M., Crooijmans, R. P., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* 4:e6524. doi: 10.1371/journal.pone.0006524

Ramsay, K., Smuts, M., and Els, H. C. (2000). Adding value to South African landrace breeds conservation through utilisation. *Anim. Genet. Resour. Inf.* 27, 9–15.

Ramsay, K. A., Reed, D. S., Bothma, A. J., and Lepen, J. M. (1994). *Profitable and Environmentally Effective Farming with Early-Domesticated Livestock in Southern Africa*. Pretoria: Department of Agriculture.

Raschetti, M., Castiglioni, B., Caroli, A., Guiatti, D., Pagnacco, G., and Chessa, S. (2013). SNP identification in swine candidate genes for meat quality. *Livest. Sci.* 155, 165–171. doi: 10.1111/age.12388

Rege, J. E. O., and Gibson, J. P. (2003). Animal genetic resources and economic development: issues in relation to economic valuation. *Ecol. Econ.* 45, 319–330. doi: 10.1016/S0921-8009(03)00087-9

Reyer, H., Oster, M., Magowan, E., Dannenberger, D., Ponsuksili, S., and Wimmers, K. (2017a). Strategies towards improved feed efficiency in pigs comprise molecular shifts in hepatic lipid and carbohydrate metabolism. *Int. J. Mol. Sci.* 18:1674. doi: 10.3390/ijms18081674

Reyer, H., Shirali, M., Ponsuksili, S., Murani, E., Varley, P. F., Jensen, J., et al. (2017b). Exploring the genetics of feed efficiency and feeding behaviour traits in a pig line highly selected for performance characteristics. *Mol. Genet. Genomics* 292, 1001–1011. doi: 10.1007/s00438-017-1325-1

Rischkowsky, B., and Pilling, D. (eds) (2007). *The State of the World's Animal Genetic Resources for Food and Agriculture*. Rome: Food and Agriculture Organization.

Rookmaaker, L. C. (1989). *The Zoological Exploration of Southern Africa 1650-1790*. Rotterdam: A.A Balkema.

Romero-Suarez, S., Shen, J., Brotto, L., Hall, T., Mo, C., Valdivia, H. H., et al. (2010). Muscle-specific inositide phosphatase (MIP/MTMR14) is reduced with age and its loss accelerates skeletal muscle aging process by altering calcium homeostasis. *J. Aging* 2, 504–513. doi: 10.18632/aging.100190

Ropka-Molik, K., Bereta, A., Żukowski, K., Tyra, M., Piórkowska, K., Żak, G., et al. (2018). Screening for candidate genes related with histological microstructure, meat quality and carcass characteristic in pig based on RNA-seq data. *Asian Australas J. Anim. Sci.* 31, 1565–1574. doi: 10.5713/ajas.17.0714

Roosenvelt, T., and Heller, E. (1915). *Life-histories of African game animals*. London: John Murray.

Ross, J. W., Hale, B. J., Gabler, N. K., Rhoads, R. P., Keating, A. F., and Baumgard, L. H. (2015). Physiological consequences of heat stress in pigs. *Anim. Prod. Sci.* 55, 1381–1390. doi: 10.1071/AN15267

Rothschild, M. F. (2010). *Association of Genetic Markers with Structural Soundness and Its Relationship to Gilt Development and Sow Longevity*. Available online at: https://www.pork.org/wp-content/uploads/2009/06/06-019-ROTHSCHILD-ISU.pdf (accessed February 20, 2017).

Rothschild, M. F., and Ruvinsky, A. (2010). *The Genetics of the Pig*. Wallingford: CABI.

Rui, L., Sun, D. X., Wang, Y., Yu, Y., Zhang, Y., Chen, H., et al. (2013). Fine mapping QTLs affecting milk production traits on BTA6 in Chinese Holstein with SNP markers. *J. Integr. Agric.* 12, 110–117. doi: 10.1016/S2095-3119(13)60211-7

Ryu, J., and Lee, C. (2016). Genetic association of marbling score with intragenic nucleotide variants at selection signals of the bovine genome. *Animal* 10, 566–570. doi: 10.1017/S1751731115002633animal

Salehi, A., Sobhani, R., Aminafshar, M., Sayyadnejhad, M. B., and Nasiri, K. (2015). Single nucleotide of *FGF2* gene in Iranian Holstein proven bulls. *Mol. Biol. Res. Commun.* 4, 57–62.

SanCristobal, M., Chevalet, C., Haley, C. S., Joosten, R., Rattink, A. P., Harlizius, M. A. M., et al. (2006). Genetic diversity within and between European pig breeds using microsatellite markers. *Anim. Genet.* 37, 189–198. doi: 10.1111/j.1365-2052.2005.01385.x

Scandura, M., Iacolina, L., and Apollonio, M. (2011). Genetic diversity in the European wild boar Sus scrofa: phylogeography, population structure and wild × domestic hybridization. *Mamm. Rev.* 41, 125–137. doi: 10.1111/j.1365-2907.2010.00182.x

Schwartz, K., Lawn, R. M., and Wade, D. P. (2000). ABC1 Gene expression and ApoA-I-Mediated cholesterol efflux regulated by LXR. *Biochem. Biophys. Res. Commun.* 274, 794–802. doi: 10.1006/bbrc.2000.3243

Siddiq, A., Gueorguiev, M., Samson, C., Hercberg, S., Heude, B., Levy-Marchal, C., et al. (2007). Single nucleotide polymorphisms in the neuropeptide Y2 receptor (NPY2R) gene and assosciation with severe obesity in French white subjects. *Diabetologia* 50, 574–584. doi: 10.1007/s00125-006-0555-2

Smith, S. P., Phillips, J. B., Johnson, M. L., Abbot, P., Capra, J. A., and Rokas, A. (2019). Genome-wide association analysis uncovers variants for reproductive variation across dog breeds and links to domestication. *Evol. Med. Public Health* 2019, 93–103. doi: 10.1093/emph/eoz015

Somers, M. J., Penzhorn, B. L., and Rasa, A. E. (1994). Home range size use and dispersal of warthogs in the Eastern Cape, South Africa. *J. Afr. Zool.* 108, 361–373.

Somers, M. J., Rasa, O. A., and Penzhorn, B. L. (1995). Group structure and social behaviour of warthogs *Phacochoerus aethiopicus*. *Acta Theriol.* 40, 257–281.

Song, J., Yang, D., Ruan, J., Zhang, J., Chen, Y. E., and Xu, J. (2017). Production of immunodeficient rabbits by multiplex embryo transfer and multiplex gene targeting. *Sci. Rep*. 7:12202. doi: 10.1038/s41598-017-12201-0

Sosa-Madrid, B. S., Santacreu, M. A., Fontanesi, L., Blasco, A., and Ibañez-Escriche, N. (2018). A genomic region on chromosome 17 has a major impact on litter size traits in rabbits. *Proc. World Congr. Genet. Appl. Livestock Product.* 11:163.

Suchocki, T., Wojdak-Maksymiec, K., and Szyda, J. (2016). Using gene networks to identify genes and pathways involved in milk production traits in Polish Holstein dairy cattle. *Czech J. Anim. Sci.* 61, 526–538. doi: 10.17221/43/2015-CJAS

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* 2, 125–141. doi: 10.1016/0040-5809(71)90011-6

Swart, H., Kotze, A., Olivier, P. A. S., and Grobler, J. P. (2010). Microsatellite-based characterization of Southern African domestic pigs (Sus scrofa domestica). *S. Afr. J. Anim. Sci.* 40, 121–132. doi: 10.4314/sajas.v40i2.57280

Sweett, H., Miglior, F., Livernois, A., Fonseca, P., Id-Lahoucine, S., Troya, E., et al. (2018). Genome-wide association study to identify genomic regions and single nucleotide polymorphisms functionally associated with bull fertility. *J. Anim. Sci.* 96, 138–139. doi: 10.1093/jas/sky404.303

Taverner, M. R., and Dunkin, A. C. (eds) (1996). *Introduction to Pig Production. Pig Production*. New York, NY: Elsevier Science.

Tizioto, P. C., Taylor, J. F., Decker, J. E., Gromboni, C. F., Mudadu, M. A., and Schnabel, R. D. (2015). Detection of quantitative trait loci for mineral content of Nelore longissimus dorsi muscle. *Genet. Sel. Evol.* 47:15. doi: 10.1186/s12711-014-0083-3

Verardo, L. L., Silva, F. F., Lopes, M. S., Madsen, O., Bastiaansen, J. W. M., Knol, E. F., et al. (2016). Revealing new candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways. *Genet. Sel. Evol.* 48:9. doi: 10.1186/s12711-016-0189-x

Vigors, S., Sweeney, T., O'Shea, C. J., Kelly, A. K., and O'Doherty, J. V. (2016). Pigs that are divergent in feed efficiency, differ in intestinal enzyme and nutrient transporter gene expression, nutrient digestibility and microbial activity. *Animal* 10, 1848–1855. doi: 10.1017/S1751731116000847

Visser, C., Lashmar, S. F., Van Marle-Köster, E., Poli, M. A., and Allain, D. (2016). Genetic diversity and population structure in South African, French and Argentinian angora goats from genome-wide SNP Data. *PLoS One* 11:e0154353. doi: 10.1371/journal.pone.0154353

Wang, H. T., Zhang, Z., Fan, M., Wang, L., and Dong, G. L. (2010). Expression of CD112 in colon carcinoma tissues and cell lines and their clinical significance. *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi* 26, 477–479.

Wang, S., Li, D., Zhu, M., Xie, R., Duan, S., Yin, Z., et al. (2019). A single nucleotide polymorphism of the porcine CXCL8 gene is associated with serum CXCL8 level. *Ital. J. Anim. Sci.* 18, 474–479. doi: 10.1080/1828051X.2018.15 39349

Wang, X., Ma, P., Liu, J., Zhang, Q., Zhang, Y., Ding, X., et al. (2015). Genome-wide association study in Chinese Holstein cows reveal two candidate gene for somatic cell score as an indicator for mastitis susceptibility. *BMC Genet.* 16:111. doi: 10.1186/s12863-015-0263-3

White, T. D., and Harris, J. M. (1977). Suid evolution and correlation of African hominid localities. *Science* 198, 13–21. doi: 10.1126/science.331477

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

Willi, Y., Van Buskirk, J., and Hoffman, A. A. (2006). Limits to the adaptive potential of small populations. *Annu. Rev. Ecol. Evol. Syst.* 37, 433–458. doi: 10.1146/annurev.ecolsys.37.091305.110145

Wright, S. (1978). *Evolution and the Genetics of Population, Variability Within and Among Natural Populations*. Chicago, IL: The University of Chicago Press.

Wu, X., Zhang, Q., Xu, S., Jin, P., Luan, P., Li, Y., et al. (2016). Differential expression of six chicken genes associated with fatness traits in a divergently selected broiler population. *Mol. Cell. Probe* 30, 1–5.

Xiang, R., Oddy, V. H., Archibald, A. L., Vercoe, P. E., and Dalrymple, B. P. (2016). Epithelial, metabolic and innate immunity transcriptomic signatures differentiating the rumen from other sheep and mammalian gastrointestinal tract tissues. *PeerJ* 4:e1762. doi: 10.7717/peerj.1762

Yang, B., Cui, L., Perez-Enciso, M., Traspov, A., Crooijmans, R. P. M. A., Zinovieva, N., et al. (2017). Genome-wide SNP data unveils the globalization of domestication pigs. *Genet. Sel. Evol.* 49:71. doi: 10.1186/s12711-017-0345-y

Yazar, S., Cuellar-Partida, G., McKnight, C. M., Quach-Thanissorn, P., Mountain, J. A., and Coroneo, M. T. (2015). Genetic and environmental factors in conjunctival UV autofluorescence. *JAMA Opthalmol.* 133, 406–412. doi: 10.1001/jamaophthalmol.2014.5627

Yoshida, G. M., Lhorente, J. P., Carvalheiro, R., and Yáñez, J. M. (2017). Bayesian genome-wide association analysis for body weight in farmed Atlantic salmon (*Salmo salar* L.). *Anim. Genet.* 48, 698–703. doi: 10.1111/age.12621

Zadik, B. J. (2005). *The Iberian Pig in Spain and the Americas at the Time of Columbus*. Master's thesis, University of California, Berkeley, 36–51.

Zeder, M. A., Emshwiller, E., Smith, B. D., and Bradley, D. G. (2006). Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.* 22, 139–155. doi: 10.1016/j.tig.2006.01.007

Zhang, W., Fan, Z., Han, E., Hou, R., Zhang, L., and Galaverni, M. (2014). Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. *PLoS Genet.* 10:e1004466. doi: 10.1371/journal.pgen.1004466

Zhang, Z., Xiao, Q., Zhang, Q. Q., Sun, H., Chen, J. C., Li, Z. C., et al. (2018). Genomic analysis reveals genes affecting distinct phenotypes among different Chinese and western pig breeds. *Sci. Rep.* 8:13352. doi: 10.1038/s41598-018-31802-x

Zhou, L., Ji, J., Peng, S., Zhang, Z., Fang, S., Li, L., et al. (2016). A GWA study reveals genetic loci for body conformation traits in Chinese Laiwu pigs and its implications for human BMI. *Mamm. Genome* 27, 610–621. doi: 10.1007/s00335-016-9657-4

Zuo, B., Xiong, Y., Yang, H., and Wang, J. (2007). Full-length cDNA, expression pattern and association analysis of the porcine FHL3 gene. *Asian Australas J. Anim. Sci.* 20, 1473–1477. doi: 10.5713/ajas.2007.1473

Zuo, B., Xiong, Y. Z., Deng, C. Y., Su, Y. H., Wang, M. G., Lei, M. G., et al. (2004). cDNA cloning, genomic structure and polymorphism of the porcine FHL3 gene. *Anim. Genet.* 35, 230–233. doi: 10.1111/j.1365-2052.2004.01119.x

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 August 2020
doi: 10.3389/fgene.2020.00880

Management of Genetic Diversity in the Era of Genomics

Theo H. E. Meuwissen[1], Anna K. Sonesson[2], Gebreyohans Gebregiwergis[1] and John A. Woolliams[3]*

[1] Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås, Norway, [2] NOFIMA, Ås, Norway, [3] The Roslin Institute and R(D)SVS, The University of Edinburgh, Edinburgh, United Kingdom

Management of genetic diversity aims to (i) maintain heterozygosity, which ameliorates inbreeding depression and loss of genetic variation at loci that may become of importance in the future; and (ii) avoid genetic drift, which prevents deleterious recessives (e.g., rare disease alleles) from drifting to high frequency, and prevents random drift of (functional) traits. In the genomics era, genomics data allow for many alternative measures of inbreeding and genomic relationships. Genomic relationships/inbreeding can be classified into (i) homozygosity/heterozygosity based (e.g., molecular kinship matrix); (ii) genetic drift-based, i.e., changes of allele frequencies; or (iii) IBD-based, i.e., SNPs are used in linkage analyses to identify IBD segments. Here, alternative measures of inbreeding/relationship were used to manage genetic diversity in genomic optimal contribution (GOC) selection schemes. Contrary to classic inbreeding theory, it was found that drift and homozygosity-based inbreeding could differ substantially in GOC schemes unless diversity management was based upon IBD. When using a homozygosity-based measure of relationship, the inbreeding management resulted in allele frequency changes toward 0.5 giving a low rate of increase in homozygosity for the panel used for management, but not for unmanaged neutral loci, at the expense of a high genetic drift. When genomic relationship matrices were based on drift, following VanRaden and as in GCTA, drift was low at the expense of a high rate of increase in homozygosity. The use of IBD-based relationship matrices for inbreeding management limited both drift and the homozygosity-based rate of inbreeding to their target values. Genetic improvement per percent of inbreeding was highest when GOC used IBD-based relationships irrespective of the inbreeding measure used. Genomic relationships based on runs of homozygosity resulted in very high initial improvement per percent of inbreeding, but also in substantial discrepancies between drift and homozygosity-based rates of inbreeding, and resulted in a drift that exceeded its target value. The discrepancy between drift and homozygosity-based rates of inbreeding was caused by a covariance between initial allele frequency and the subsequent change in frequency, which becomes stronger when using data from whole genome sequence.

Keywords: inbreeding, genetic drift, optimum contribution selection, genetic diversity, genomic relationships, genetic gain

OPEN ACCESS

Edited by:
Maria Saura,
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Spain

Reviewed by:
Jesús Fernández,
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Spain
Yoshitaka Nagamine,
Nihon University, Japan
Piter Bijma,
Wageningen University and Research, Netherlands

****Correspondence:***
Theo H. E. Meuwissen
theo.meuwissen@nmbu.no

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *16 May 2019*
Accepted: *17 July 2020*
Published: *13 August 2020*

Citation:
Meuwissen THE, Sonesson AK, Gebregiwergis G and Woolliams JA (2020) Management of Genetic Diversity in the Era of Genomics. Front. Genet. 11:880. doi: 10.3389/fgene.2020.00880

BACKGROUND

Management of genetic diversity is usually directed at maintaining the diversity that was present in some population, which serves as a reference point against which diversity in the future is compared. This reference population may be some population in the past or the current population. In the absence of genomic data, the accumulated change in diversity was predicted to be a loss, and could only be described by inbreeding coefficients (F) based on pedigree data. These coefficients are the expectations of the loss in genetic variance relative to the reference population in which all alleles are assumed to be drawn at random with replacement, i.e., the classical base population. This description as a loss of variance is strictly for additive traits, but individual allele frequency at a locus among individuals (i.e., 0, ½, 1) is an additive trait. In this perspective, the management of genetic diversity comes down to the management of inbreeding, in particular controlling the rate of inbreeding (ΔF), or, equivalently, the effective population size: $N_e = 1/(2\Delta F)$ (Falconer and Mackay, 1996).

Optimal management of inbreeding in breeding schemes is achieved by optimal contribution (OC) selection (Meuwissen, 1997; Woolliams et al., 2015) that, by construction, maximizes the genetic gain made for a given rate of inbreeding. In the era of genomics, Sonesson et al. (2012) concluded that genomic selection requires genomic control of inbreeding, i.e., genomic optimal contribution selection (GOC). With OC, the management of diversity within the population uses the form $\frac{1}{2}c'Ac$ where **A** is wright's numerator relationship matrix and **c** is a set of fractional contributions of candidates to the next generation, and with GOC a genomic relationship matrix **G** replaces **A**. This has direct correspondence with the substantial literature on the use of similarity matrices and the fractional contributions of species as measures of species diversity (e.g., Leinster and Cobbold, 2012). The similarity matrices in OC use the idea of relationships, which are the scaled (co)variances of breeding values between all pairs of individuals in a population past and present, which links to the wider canon of genetic theory.

In the pre-genomics era, relationships were based on pedigree and pedigree-based coefficients of kinship describing the probability of identity-by-descent (IBD) at neutral loci that are unlinked to any loci under selection. Within this subset of loci, IBD results in a redistribution of genotype frequencies away from Hardy-Weinberg proportions toward homozygosity by $p_0^2(1-F)+p_0F$, $2p_0\left(1-p_0\right)(1-F)$, and $(1-p_0)^2(1-F)+\left(1-p_0\right)F$ for the genotypes AA, Aa and aa, respectively, where p_0 is the original frequency of the A allele (Falconer and Mackay, 1996). This redistribution of genotype frequencies links the changes of heterozygosity [expected to reduce by a factor (1–F)], the within line genetic variance [also reducing by (1–F)], and the genetic drift variance of allele frequencies [$p_0(1–p_0)$F] to the inbreeding coefficient describing the IBD of sampled alleles. These expected changes do not hold for loci linked to the causal variants of complex traits (QTL), where allele frequencies and genotype frequencies may change non-randomly, and cannot be explained by IBD predicted by pedigree alone.

When defining inbreeding as the correlation between uniting gametes, Wright (1922) assumed the infinitesimal model, which implies infinitesimal selection pressures with random changes in allele frequency. However, the genome is of finite size, and for complex traits with many QTL selection pressures will extend to neutral loci in linkage disequilibrium (LD) across the genome, and these associations to loci under selection result in non-random changes of allele frequencies. This is particularly the case for genomic selection schemes, where marker panels are large, but not infinitely large, dense and genome-wide, and designed to be in LD with all QTL, and where selection is directly for the markers included in the panel. In this setting unlinked neutral loci are likely to be rare, so the classical theory appears redundant.

Despite the apparent loss of a unifying paradigm, genomics opens up a choice of tools that could be used to describe genetic diversity that is wider in scope than the classical genetic variance and inbreeding. For example, tools based on genomic relationships (VanRaden, 2008), runs of homozygosity (de Cara et al., 2013; Luan et al., 2014; Rodríguez-Ramilo et al., 2015), and linkage analysis (Fernando and Grossman, 1989; Meuwissen et al., 2011). Some genomic measures may be better suited for some purposes than others, and so the question arises of what is the purpose of the management of diversity in breeding schemes in addition to what tools to use. Furthermore, when considering tools for genomic inbreeding, there is a need to distinguish which aspect of inbreeding they depict (IBD, heterozygosity/homozygosity, or genetic drift), since in (genomic) selection schemes their expectations may differ from those derived from random allele frequency changes resulting in the genotype frequencies $p_0^2(1-F)+Fp_0$, $2p_0\left(1-p_0\right)(1-F)$, and $(1-p_0)^2(1-F)+F\left(1-p_0\right)$.

Most molecular genetic measures of inbreeding are based on the allelic identity of marker loci, and do not directly separate IBD from Identity-By-State (IBS). Genomic relationship matrices which are variants of VanRaden (2008) compensate for this by measuring squared changes in allele frequency relative to a set of reference frequencies. For the purposes of managing changes in diversity relative to the reference population these frequencies would be those relevant to this base generation (Sonesson et al., 2012), although often the frequencies in the current "generation" are used (Powell et al., 2010), or simply the subset of the population for which the genomic data is available; see Legarra (2016) for further discussion on these issues. Providing the base generation is used to define the reference frequencies at neutral unlinked loci ($p_{0,k}$ for locus k), the expectation of $\mathbf{G}_{\mathrm{VR2}}$ (Method 2; VanRaden, 2008) is **A**, with all loci equally weighted after standardization using the base generation frequencies. In comparison, $\mathbf{G}_{VR1}$ (Method 1) can be viewed as simply re-weighting the loci by $2p_{0,k}(1-p_{0,k})$: i.e., for a single locus, $\mathbf{G}_{VR1}$ and $\mathbf{G}_{\mathrm{VR2}}$ yield identical relationship estimates, and extending to many loci $\mathbf{G}_{VR2}$ uses the simple mean of the single locus estimates whereas $\mathbf{G}_{VR1}$ uses the weighted mean with $2p_{0,k}(1-p_{0,k})$ as the weights. Extending the argument of Woolliams et al. (2015) for $\mathbf{G}_{VR1}$, since $\mathbf{G}_{\mathrm{VR2}}$ is based on the squares of standardized allele frequency changes, and the management of diversity using $\mathbf{G}_{\mathrm{VR2}}$ will constrain these squared standardized changes; this measurement of inbreeding will be denoted as F_{drift} [see Eq. (1B)

in Methods section for a more precise definition]. When using 0.5 as the base frequency for all loci, as sometimes proposed, the relationship matrix $G_{VR_{0.5}}$ is proportional to homozygosity and molecular coancestry (Toro et al., 2014). Hence, $G_{VR_{0.5}}$ may be used to measure homozygosity-based inbreeding, F_{hom}, and the loss of heterozygosity ($1-F_{\text{hom}}$).

The use of a genomic relationship matrix, $\mathbf{G}_{\text{LA}}$, based on linkage analysis for inbreeding management was suggested and studied by Toro et al. (1998), Wang (2001), Pong-Wong and Woolliams (2007), Fernandez et al. (2005), and Villanueva et al. (2005). Here the inheritance of the marker alleles is used to determine probabilities of having inheriting the maternal or paternal allele from a parent at the marker loci instead of assuming 50/50 inheritance probabilities as in **A**. $\mathbf{G}_{\text{LA}}$ thus requires pedigree and marker information, and IBD relationships are relative to the (assumed) unrelated and non-inbred base population as in **A**. In this way IBD is evaluated directly by $\mathbf{G}_{\text{LA}}$, and is not simply an expectation for neutral unlinked loci as described above for $\mathbf{G}_{\text{VR2}}$. If two (base) individuals are unrelated in **A** then they are unrelated in $\mathbf{G}_{\text{LA}}$, whereas the other measures also estimate (non-zero) relationships for base population individuals. The marker data accounts for Mendelian segregation which may deviate from 50/50 probabilities through any linkage drag from loci under selection, or selective advantage. $\mathbf{G}_{\text{LA}}$ can be constructed by a tabular method, similar to that for the pedigree based relationship matrix (Fernando and Grossman, 1989), and software for the simultaneous linkage analysis of an entire chromosome is available (e.g., LDMIP (Linkage Disequilibrium Multilocus Iterative Peeling); Meuwissen and Goddard, 2010). $\mathbf{G}_{\text{LA}}$ is a tool that specifically describes IBD across the genome, hence we will denote this IBD based estimate of inbreeding as F_{IBD}.

A run of homozygosity (ROH) is an uninterrupted sequence of homozygous markers (McQuillan et al., 2008). The exact definition of a ROH differs among studies as a number of ancillary constraints are added related to the minimum length of a ROH measured in markers and/or cM, minimum marker density, and in some cases an allowance for some heterozygous genotypes arising from genotyping errors. The idea is that a run of homozygous markers indicates an IBD segment, since it is unlikely that many consecutive homozygous markers are IBS by chance alone. The total length of ROH relative to the total genome length provides an estimate of F_{IBD} from the DNA itself, and this estimate will be denoted F_{ROH}. The reference population for F_{ROH} is unclear, although by varying the constraint on the length of the ROHs the emphasis can be changed from old inbreeding, with short ROHs, to young inbreeding, with long ROHs (Keller et al., 2011). F_{ROH} may miss some relevant inbreeding since IBD segments shorter than the minimum length are neglected. On the one hand, F_{ROH} is an IBD based measure of inbreeding, as it attempts to identify IBD segments (especially when ROHs are long), but on the other hand it is a homozygosity based measure of inbreeding since it is actually based on the homozygosity of haplotypes (especially when ROHs are short). However, F_{ROH} is a measure of inbreeding in a single individual and is unsuitable for a measure of IBD within the population as a whole. Therefore integration of ROH into a GOC framework requires a pairwise measurement to form a similarity matrix, $\mathbf{G}_{\text{ROH}}$ (de Cara et al., 2013).

The aim of this study is to: (i) re-examine the goals of the management of genetic diversity in breeding schemes, and the molecular genetic parameters that may be incorporated into these goals; and (ii) compare alternative genomic- and pedigree-based measures of inbreeding and relationships for addressing the goals. In doing so the different tools discussed above and some novel variants will be compared for their ability to generate gain in breeding schemes while measures of inbreeding are constrained. Finally, conclusions are made with respect to the practical implementation of these tools for managing diversity and how the outcomes will depend on whether whole genome sequence (WGS) data is considered or marker panels.

MATERIALS AND METHODS

The Goals of the Management of Genetic Diversity

Managed populations, such as livestock, will generally have many desirable characteristics (related to production, reproduction, disease resistance, etc.). Some of these characteristics are to be improved (the breeding goal traits), without jeopardizing the others. The latter is the aim of the management of inbreeding. Specifically, breeding programs aim to change allele frequencies at the QTL in the desired direction. This ultimately results in loss of variation at the QTL as fixation approaches, but providing these changes are in the right direction this loss of variation is not a problem. However, genetic drift from our reference population and loss of variation at loci that are neutral for the selection goal are to be avoided for the following reasons. Firstly, to alleviate the risk of inbreeding depression through decreased heterozygosity, particularly for traits that are not under artificial selection but are needed for the healthy functioning of the animals. Secondly, deleterious recessive alleles may drift to high frequencies, and occur more frequently in their deleterious or lethal homozygous form; although mentioned separately this is a specific manifestation of inbreeding depression. In the genomics era, deleterious recessives may be identified and mapped (Charlier et al., 2008), and if achieved recessive mutations may be selected against (at the cost of selection pressures), or potentially gene-edited. Nonetheless, simultaneous selection against many genetic defects diverts substantial selection pressures away from other traits in the breeding goal. Thirdly, loss of variation arising from selection sweeps for the current goal may erase variation for traits that are currently not of interest but may be valued in the future and so limit the future selection opportunities. Fourthly, genetic drift in the sense of random changes of allele frequencies, and thus random changes of trait values, which may be deleterious. This encompasses both the traits outside the current breeding goal and within it, where drift is observed as variability in the selection response. Moreover, large random changes in allele frequency may disrupt positive additive-by-additive interactions between QTL which have occurred due to many generations of natural and/or artificial selection (similar to recombination losses in crossbreeding; Kinghorn, 1980). In

addition, random allele frequency changes may result in the loss of rare alleles, which implies a permanent loss of variation.

Measures for Management of Inbreeding

Whilst genomics offers molecular measures for direct monitoring, most obviously heterozygosity and frequency changes measured from a panel of anonymous markers, the strategy for management of these diverse problems using genomics does not follow directly. For example, increasing heterozygosity *per se*, achieved by moving allele frequencies of marker loci toward ½ is not solely beneficial, as while potentially ameliorating the aforementioned problems 1 and 3 it is deleterious for problems 2 and 4. Both these empirical measures of heterozygosity and the change of frequencies from drift can be considered to be measures of inbreeding and diversity. Wright (1922) states that a natural inbreeding coefficient moves between 0 and 1 as heterozygosity with random mating moves between its initial state and 0: therefore, if a locus k has initial frequency p_0 and current frequency $p_{t,k}$ then a measure of inbreeding is $1-(H_{t,k}/H_{0,k}) = 1-[2p_{t,k}(1-p_{t,k})]/[2p_{0,k}(1-p_{0,k})]$, which can be generalized by averaging loci to obtain F_{hom}, i.e.,

$$F_{\text{hom}} = 1 - \sum_{\text{loci } k} \frac{2p_{t,k}\left(1-p_{t,k}\right)}{2p_{0,k}\left(1-p_{0,k}\right)}/N_{\text{SNP}} \qquad (1A)$$

where N_{SNP} is the total number of loci. F_{hom} can be negative when heterozygosity increases due to allele frequencies moving toward 0.5. Similarly, drift can be measured as $\delta p_{t,k}^2 = (p_{t,k}-p_{0,k})^2$, scaled by the expected value for complete random inbreeding, i.e., $\delta p_{t,k}^2/[p_{0,k}(1-p_{0,k})]$, and similarly averaged over loci to obtain F_{drift}, i.e.,

$$F_{\text{drift}} = \sum_{\text{loci } k} \frac{\delta p_{t,k}^2}{p_{0,k}\left(1-p_{0,k}\right)}/N_{\text{SNP}} \qquad (1B)$$

and which is never negative. F_{drift} is similar to the definition of F_{ST} (Holsinger and Weir, 2009), which is here applied to a single population over time instead of a sample of populations, and it is this empirical measure that is being directly addressed when using G_{VR2} .

For locus k in the set of neutral loci with frequency $p_{0,k}$ in the base population and frequency $p_{t,k} = p_{0,k} + \delta p_{t,k}$ in generation t, twice the frequency in generation t is $2p_{t,k}^2 + H_{t,k} = 2(p_0 + \delta p_{t,k})$, where $H_{t,k} = 2(p_0 + \delta p_{t,k})(1 - p_0 - \delta p_{t,k})$, which holds for all loci assuming random mating. With a sufficiently large subset of neutral loci with the same base frequency p_0 if $E[\delta p_{t,k}|p_0] = 0$ then taking expectations over this subset $2E[p_{t,k}^2] + E[H_{t,k}] = 2p_0$ and so $2(E[p_{t,k}^2] - p_0^2) + E[H_{t,k}] = 2p_0(1-p_0)$. The first term is $2var(p_{t,k})$ and the second is H_t and dividing through by $2p_0(1-p_0)$ gives

$$var\left(p_{t,k}\right) / \left[p_0\left(1-p_0\right)\right] = 1 - H_{t,k}/H_0 \quad \Rightarrow \quad F_{\text{drift}} = F_{\text{hom}} \qquad (2)$$

Therefore if $E[\delta p_{t,k}|p_0] = 0$ over the range $0 < p_0 < 1$, there is an equivalence of F_{drift} with F_{hom} irrespective of initial frequency, p_0 (Falconer and Mackay, 1996): i.e., drift- and homozygosity-based inbreeding are expected to be the same if allele frequency changes are on average 0 irrespective of the initial frequency.

Using a form of GOC related to G_{VR1} (see Discussion), de Beukelaer et al. (2017) explore the management of diversity and derived the consequences for the rate of homozygosity, $2(\delta p_{t,k}^2 + 2\delta p_{t,k}\left(p_0 - \frac{1}{2}\right))/H_{t,k}$. They suggested (supported by results below) that the term $\delta p_{t,k}(p_0 - \frac{1}{2})$, which represents a covariance between allele frequency change $\delta p_{t,k}$ and initial frequency $p_{0,k}$ across the loci k, may be non-zero. Consequently, $E[\delta p_{t,k}|p_0] \neq 0$, and Equation [2] will no longer hold, and $F_{\text{drift}} \neq F_{\text{hom}}$. **Supplementary Information 1** shows that *any* deviation from Equation [2] for a general set of loci for which $E[\delta p_{t,k}] = 0$ over the set, not necessarily with the same initial frequency, must be explained by a covariance between allele frequency changes and the original frequency $\text{cov}(\delta p_{t,k}; p_{0,k})$ and shows:

$$F_{\text{hom}} - F_{\text{drift}} = 2\text{cov}(\delta p_{t,k}/\sqrt{p_{0,k}(1-p_{0,k})}; \\ (p_{0,k^{-1/2}})/\sqrt{p_{0,k}(1-p_{0,k})}) \qquad (3)$$

i.e., if there is covariance between initial allele frequencies and frequency changes, homozygosity and drift based inbreeding are no longer equal. Therefore this covariance will be important in determining the impact of genomic management, which aims to manage both the increase of homozygosity and genetic drift.

Supplementary Information 1 explores why completely random selection of parents (i.e., with no management) generates no covariance and how different broad management goals for diversity may generate a covariances of different signs. In particular, with completely random selection, most markers drift to the nearest extreme with the smaller change in frequency, but a minority will move to the opposite extreme resulting in the larger frequency change, giving a net result of no covariance. The consequence of using GOC based on G_{VR2} is that the latter large allele frequency changes are penalized more heavily, since they add as $\delta p_{t,k}^2$ to the elements of G_{VR2} and consequently to $\frac{1}{2}c'Gc$. Hence, the hypothesis is tested below that G_{VR2} emphasizes the movement of MAF toward 0, and more generally allele frequencies move away from intermediate values toward the nearest extreme, resulting in $cov(\delta p_{t,k}; p_{0,k}) > 0$ and $var\left(p_{t,k}\right)/[p_0(1-p_0)] + E[H_{t,k}/H_{0,k}] < 1$, contrary to expectations in Eq. (2).

Conversely if $G_{0.5}$ is used in GOC then there will be pressure to move allele-frequencies toward 0.5 resulting in increasing heterozygosity (Li and Horvitz, 1953). **Supplementary Information 1** shows that this results in $cov(\delta p_{t,k}; p_{0,k}) < 0$, and thus $F_{\text{hom}} < 0$, and $F_{\text{drift}} > 0$, and $var\left(p_{t,k}\right)/[p_0(1-p_0)] + E[H_{t,k}/H_{0,k}] > 1$, again contrary to expectations in Eq. (2). Furthermore the implication of these considerations is that the covariance $cov(\delta p_{t,k}; p_{0,k})$ is a property of the active management of diversity using squared frequency changes as in G_{VR2} (or G_{VR1}) and not as a consequence of directional selection. This hypothesis was tested below in two ways: firstly by combining the management of diversity using G_{VR2} with randomly generated EBVs, and secondly by using a panel of markers for managing

diversity that is distinct from the panel used for estimating GEBVs for genomic selection.

The term $\delta p_{t,k}^2/[p_{0,k}(1-p_{0,k})]$ appearing in F_{drift} can be viewed as an approximation to the squared total intensity (i^2) applied to the marker, where $i \approx \delta p_{t,k}/\left[p_{0,k}\left(1-p_{0,k}\right)\right]$. The approximation arises because the total selection intensity applied to a marker is not linear with frequency (see Liu and Woolliams, 2010). For example, after the initial generation, the intensity applied to alleles moved toward ½ is overestimated, since the denominator of i increases over time, which reduces the actual intensity applied. The opposite holds for those alleles moved toward the nearest extreme. Therefore a further hypothesis is that a relationship matrix built upon i^2, $G_{i(p)}$, rather than $\delta p_{t,k}^2$ may remove the covariance of the change in frequency with the initial frequency that is generated using G_{VR2}. More details on this and the calculation of $G_{i(p)}$ are given in **Supplementary Information 2**.

In classical theory, the equivalence of F_{drift} withF_{hom} under random mating is an outcome of considering IBD, and management by IBD. The genomic relationship matrices based on allele frequency changes or functions of these changes no longer consider IBD as they only consider IBS. **Supplementary Information 3** considers the IBD properties of the linkage analysis relationship matrix G_{LA} which is derived from the markers. Considering the management of diversity over generations when using G_{LA}, the conclusion of **Supplementary Information 3** is that $\delta p_{t,k}$ will now be determined by the properties of the base population and not through linkage disequilibrium generated in the course of the selection process. Therefore, the covariance between the change in frequency and its initial value is potentially avoided. This leads to a further hypothesis tested below that if G_{LA} replaces G_{VR2} in GOC then $F_{\text{drift}} = F_{\text{hom}}$ and $var\left(p_{t,k}\right)/[p_0(1-p_0)] + E[H_{t,k}/H_{0,k}] = 1$, as expected in Eq. (2); i.e., consideration of IBD restores the equivalence of F_{drift} and F_{hom} for a set of neutral markers. If A or a ROH-based G_{ROH} replaces G_{LA} the same hypothesis may be advanced given their focus on approximating IBD, however, both are approximations to the true genomic IBD that is tracked by G_{LA} and so the equivalence may only be approximate.

In summary, there are a range of hypotheses to be tested on three categories of relationship matrix: those based on drift, changes in allele frequency or functions of them $\left(G_{VR1}, G_{VR2}, \text{and } G_{i(p)}\right)$; those based on homozygosity exemplified by $G_{0.5}$; and those based on IBD (G_{LA} and **A**). A relationship matrix based on ROH, G_{ROH}, is a hybrid of the latter two, targeting IBD by measuring homozygosity of haplotypes.

Breeding Structure and Genomic Architecture

A computer simulation study was conducted to compare these alternative GOC methods. The simulations mimicked a breeding scheme using sib-testing, such as those used for disease challenges in fish breeding, which is similar to Sonesson et al. (2012). The scheme had a nucleus where selection of candidates was entirely based on their genomic data and performance recording was solely on the full-sibs of the selection candidates which were also genotyped. This scheme may be considered extreme in the sense that the candidates themselves have no performance records, and is practiced in aquaculture to prevent disease infections within the breeding population. There were 2000 young fish per generation, and every full-sib family was split in two: half of the sibs became selection candidates and the other half test-sibs. The actual number of families and their size depended on the optimal contributions of the parents.

The genome consisted of 10 chromosomes of size 1 Morgan. Base population genomes were simulated for a population of an effective size of N_e = 100 for 400 (=$4N_e$) generations with SNP mutations occurring at a rate of 10^{-8} per base pair per generation using the infinite-sites model. This resulted in WGS data for base population genomes that were in mutation-drift-linkage disequilibrium balance. The historical population size was chosen to equal the effective population size targeted for the breeding schemes and so avoid any effect of a sudden large change in effective population size. This resulted in 33,129 segregating SNP loci, which is relatively small in number due to the small effective size of 100. From these loci $N_{\text{SNP}} = 7000$ were randomly sampled as marker loci for use in obtaining GEBV by genomic selection (Panel M); another distinct sample of 7000 loci were randomly sampled as additive QTL, which obtained an allelic effect sampled from the Normal distribution (Panel Q); and a further distinct sample of 7000 SNP loci were randomly sampled to act as "neutral loci" (Panel N), which were used to assess allele-frequency changes and loss of heterozygosity at neutral (anonymous) WGS loci, not involved in either genomic prediction or diversity management. In the majority of schemes Panel M was used for constructing genomic relationship matrices for both obtaining EBVs and diversity management. However, to test whether the non-neutrality of the SNPs used for genomic prediction interfered with their simultaneous use for diversity management, a further distinct panel of 7000 randomly picked loci (Panel D) was used for diversity management in some schemes.

True breeding values were obtained by summing the effects of the QTL alleles across the loci in Panel Q, before scaling them such that the total genetic variance was $\sigma_g^2 = 1$ in the base population. Phenotypes were obtained by adding a randomly sampled environmental effect with variance $\sigma_e^2 = 1.5$, resulting in a heritability of 0.4. After the initial 400 unselected generations to simulate a base population (t = 0), the breeding schemes described below were run for 20 generations, of which the first generation comprised random selection in order to create an initial sib-family structure.

Genomic Estimates of Breeding Values

GEBV ($\widehat{g}$) were obtained by the SNP-BLUP method (Meuwissen et al., 2001) where BLUP estimates of SNP effects were obtained from random regression on the SNP genotypes of Panel M coded as $X_{ik} = -2p_{0,k}/\sqrt{[2p_{0,k}(1-p_{0,k})]}$, $(1-2p_{0,k})/\sqrt{[2p_{0,k}(1-p_{0,k})]}$, or $(2-2p_{0,k})/\sqrt{[2p_{0,k}(1-p_{0,k})]}$ for homozygote, heterozygote, and alternative homozygote genotypes, respectively, of the kth SNP of animal i, and $p_{0,k}$ is the allele frequency of a randomly chosen

reference allele of the kth SNP in generation 0. The model for the BLUP estimation of the SNP effects was:

$$y = 1\mu + Xb + e$$

where $\boldsymbol{y}$ is a vector of records; μ is the overall mean; $\boldsymbol{X}$ is a matrix of genotype codes as described above; $\boldsymbol{b}$ is a vector of random SNP effects [*a priori*, $b \sim MVN(0, \sigma_g^2 N_{SNP}^{-1} I)$], and $\boldsymbol{e}$ is a vector of random residuals [*a priori* $e \sim N(0, \sigma_e^2 I)$]. GEBV were obtained as $\hat{g} = X\hat{b}$ where $\hat{b}$ denotes the BLUP estimates of the SNP effects. This model is often implemented in the form of GBLUP using VanRaden (2008) Model 2, which assumes that all loci explain an equal proportion of the genetic variance. When simulating true breeding values, variances of allelic effects were equal across the loci, which implies that the high-MAF QTL explain more variance than the low-MAF QTL. Hence, there is a discrepancy between the simulation model and that used for analysis. However, such discrepancies always occur with real data. To separate the effects of selection and inbreeding management, one of the schemes described below randomly sampled GEBVs from a Normal distribution each generation.

Assessing the Rates of Inbreeding at Neutral Loci

F_{hom} and F_{drift} were calculated for each scheme, and since discrepancies were anticipated (**Supplementary Information 1**) ΔF was also calculated from both heterozygosity and drift to give ΔF_{hom} and ΔF_{drift}. The calculations described below were done for all schemes with Panel N which were both functionally neutral in not influencing the breeding goal traits, and algorithmically neutral in not being involved in the breeding value prediction. Calculations were repeated for Panel M, and Panel D when used.

Heterozygosity

Calculation was based upon classical models where for generation t $(\Sigma_{\text{loci}\,k} H_{t,k}/H_{0,k})/N_{\text{SNP}} = 1 - F_{\text{hom}} = (1 - \Delta F)^t$ where ΔF is the rate of inbreeding, and N_{SNP} the number of loci in the panel. A log transformation yields a linear relationship $\log(\Sigma_{\text{loci}\,k} H_{t,k}/H_{0,k}) - \log(N_{\text{SNP}}) = t\log(1 - \Delta F) \approx -t\Delta F$, where the approximation holds for small ΔF when using natural logarithms. This regression was calculated and provided both a test of constant ΔF_{hom} and an estimate of ΔF_{hom} from $(-1) \times$ slope of the regression.

Drift

At time t, F_{drift} was calculated as $\Sigma_{\text{loci}\,k}(p_{t,k} - p_{0,k})^2/[p_{0,k}(1 - p_{0,k})]$. Analogously with heterozygosity, classical theory was followed by taking logs of $(1 - F_{\text{drift}})$ with ΔF_{drift} estimated by $-1 \times$ slope from the regression on t.

Optimum Contribution Selection Methods

In optimal contribution selection, the rate of inbreeding is constrained by constraining the increase of the group coancestry of the selected parents, $\overline{G} = \frac{1}{2} c' G c$, where G denotes the relationship matrix of interest for managing diversity among the selection candidates, and $\mathbf{c}$ denotes a vector of contributions of the selection candidates to the next generation, which is proportional to their numbers of offspring. Therefore the group coancestry is the average relationship among all pairs of the parents, including self-pairings, weighted by the fraction of offspring from the pair assuming completely random mating. Furthermore, the genetic level of the selected animals, $\bar{g} = c'\hat{g}$, is maximized weighted by their number of offspring. Hence, the optimisation is as follows:

$$\begin{aligned} \text{maximize} \quad & \bar{\mathbf{g}} = \text{c'}\hat{\mathbf{g}} \text{ by varying } \mathbf{c} \\ \text{with constraints :} \quad & K = \frac{1}{2}\text{c'Gc} \\ & \Sigma_{j \text{ males}} \; c_j = \frac{1}{2} \\ & \Sigma_{j \text{ females}} \; c_j = \frac{1}{2} \\ & c_j \geq 0 \text{ for all } j. \end{aligned}$$

A number of relationship matrices were investigated for managing the diversity: (i) the pedigree-based relationship matrix $\mathbf{A}$; (ii) the genomic relationship matrix $\mathbf{G}_{\text{VR2}} = XX'/N_{\text{SNP}}$ (VanRaden, 2008; Model 2) constructed using Panel M; (iii) the genomic relationship matrix $G_{VR1} = ZZ'/\Sigma_{\text{loci}\,k} H_{0,k}$ (VanRaden, 2008; Model 1) constructed using SNP Panel M where $Z_{ij} = \left(-2p_{0j}\right)$, $(1 - 2p_{0j})$, or $\left(2 - 2p_{0j}\right)$; (iv) $G_{0.5}$, a homozygosity based matrix of relationships, since its elements (i,j) are proportional to the expected homozygosity of progeny of animals i and j (Toro et al., 2014); (v) G_{LA} constructed from Panel M using linkage analysis (Fernando and Grossman, 1989; Meuwissen et al., 2011); (vi) a novel relationship matrix $G_{i(p)}$ constructed from squared total applied intensities using Panel M (see **Supplementary Information 2**); (vii) the genomic relationship matrix G_{ROH} based on ROH assessed using Panel M following the method of de Cara et al. (2013) (see **Supplementary Information 2**); (viii) a genomic relationship matrix G_{VR2} constructed using Panel D instead of M. In this replicated simulation study, the calculation of G_{LA} by LDMIP (Meuwissen and Goddard, 2010) was computationally too demanding and instead, a haplotype-based approach was adopted as an approximation (see **Supplementary Information 2**).

Implementation of Selection Procedures

The selection schemes simulated will be denoted by the relationship matrix used in GOC and the panel of markers used for SNP-BLUP and building the relationship matrix. The panel for SNP-BLUP was either "M", or "∼" when using randomly generated GEBV. The latter implements a scheme without directional selection, and tests whether observed results are due to selection or due to diversity management. The panel for management of inbreeding was either "M," "D," or "∼" when using $\mathbf{A}$ which required no marker panel. Therefore a total of 9 schemes contribute to the results presented: 6 of which are of the form $\mathbf{G}$(M,M) where $\mathbf{G}$ is either $\mathbf{G}_{VR1}$, $\mathbf{G}_{\text{VR2}}$, $\mathbf{G}_{0.5}$, $\mathbf{G}_{\text{LA}}$, $\mathbf{G}_{i(p)}$, and $\mathbf{G}_{\text{ROH}}$; with the remaining three being $\mathbf{A}$(M,∼), $\mathbf{G}_{\text{VR2}}$(M,D), and

TABLE 1 | The relationship matrices and marker panels that were used for the alternative breeding schemes.

	Marker Panel[1]		F-management	
Scheme[2]	EBV-estimation	F-management	Matrix[3]	Type of measure
G_{VR2}(M,M)	M	M	G_{VR2}	Drift
G_{VR2}(M,D)	M	D	G_{VR2}	Drift
G_{VR2}(∼,M)	∼	M	G_{VR2}	Drift
G_{VR1}(M,M)	M	M	G_{VR1}	Drift
$G_{i(p)}$(M,M)	M	M	$G_{i(p)}$	Drift
$G_{0.5}$(M,M)	M	M	$G_{0.5}$	Homoz.
G_{ROH}(M,M)	M	M	G_{ROH}	Homoz.
G_{LA}(M,M)	M	M	G_{LA}	IBD
A(M,∼)	M	∼	A	IBD

[1]M = regular marker panel used for selection (and management); D = an extra marker panel solely used for inbreeding management (if used at all); ∼ = no markers needed for management / selection (∼ for selection implies random selection). [2]Breeding schemes are denoted by G(P_{EBV},P_F) where G denotes the relationship matrix used for inbreeding management (all schemes used $\mathbf{G}_{VR2}$ for EBV estimation); P_{EBV} denotes the marker panel used for EBV estimation; and P_F denotes the marker panel used for inbreeding management. [3]Abbreviations and calculations of the relationship matrices are explained in the main text.

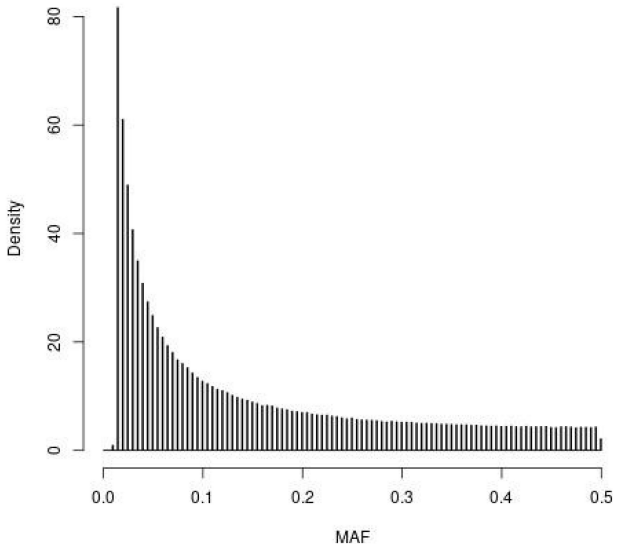


FIGURE 1 | Histogram of the minor allele frequencies (MAF) of the SNPs in the whole genome sequence of the founder population (t = 0) observed in the simulations following 4000 generations of mutation and random selection.

$\mathbf{G}_{VR2}$(∼,M), where the first symbol in parentheses refers to EBV estimation and the second to diversity management. The schemes are summarized in **Table 1**.

For all schemes the target ΔF was set via the parameter K to 0.005 / generation, so the target effective population size was 100. Therefore the group coancestry of the parents was set in generation t to $K_t = K_{t-1} + 0.005\,(1 - K_{t-1})$, where $K_0 = {}^1/_2\bar{G}$ and $\bar{G}$ denotes the average relationship of all candidates in generation 1 (the first generation with GOC selection). Each scheme was replicated 100 times by generating a new base population as described above. Simulation errors were reduced by simulating all alternative breeding schemes on each replicate of the initial generations, using the same Panels M, Q, N, and D, and the same effects for the QTLs. Each generation had random mating among males and females with mating proportions guided by the optimum contributions **c**.

$\mathbf{G}_{LA}$ and **A** are mathematically guaranteed to be positive definite, and $\mathbf{G}_{VR1}$, $\mathbf{G}_{VR2}$, $\mathbf{G}_{0.5}$, and $\mathbf{G}_{i(p)}$ are guaranteed to be positive semi-definite, i.e., all eigenvalues $\lambda_i \geq 0$, as they are the cross-product of SNP genotype matrices (**X** or **Z**) with one eigenvalue of zero due to the centring of the genotypes. For the semi-definite matrices a small value (α = 0.01) was added to their leading diagonal to make them invertible, and positive definite to permit the use of the optimal contribution algorithm of Meuwissen (1997). In contrast, $\mathbf{G}_{ROH}$ is not guaranteed to be semi-positive definite since its elements are calculated one by one, and large negative eigenvalues for $\mathbf{G}_{ROH}$ were observed empirically (results not shown). When using a general matrix inversion routine the achieved ΔF were much larger than 0.005/generation. Hence, $\mathbf{G}_{ROH}$ was made positive definite by adding substantial values of α to its diagonals, chosen by trial and error. Starting from an initial value of α = 0.05, positive definiteness was tested by inversion using Cholesky decomposition, and if it failed then α was doubled if α < 1 or increased by 1 otherwise, until inversion was successful.

RESULTS

SNPs

The distribution of MAF for the SNPs in the WGS of the founder population (t = 0) observed in the simulations is depicted in **Figure 1**. The four SNP panels, i.e., M, the SNP-BLUP panel, N, the neutral marker panel, Q, the QTL panel, and D, a second marker panel for genetic diversity management, are random samples from the SNPs depicted in **Figure 1**. The MAF distribution is typical for that of whole genome sequence data with very many SNPs with rare alleles and relatively few SNPs with intermediate allele frequencies.

Equivalence of F_{drift} and F_{hom}

Table 2 shows for the alternative breeding schemes the drift- and homozygosity-based rates of inbreeding, together with the deviations F_{hom}–F_{drift} in generation 20. For classical inbreeding theory the expectation is that F_{hom} = F_{drift} = 0.095 for random mating. However, with two sexes there will be deviations which depend on the number of mating parents which are shown in **Figure 2** and were approximately equally divided between males and females each generation. This has an impact in decreasing F_{hom} at generation 20 below random mating expectations by approximately 1/(2T) where T is the total number of parents following Robertson (1965). Therefore at generation 20, there is a classical expectation for F_{drift} to exceed F_{hom} by ∼0.001 for schemes $\mathbf{G}_{ROH}$(M,M) and **A**(M,∼), through ∼0.005 for $\mathbf{G}_{LA}$(M,M) to ∼0.01 for $\mathbf{G}_{VR2}$(M,M).

TABLE 2 | Rates of increase of homozygosity (ΔF_{hom}), drift (ΔF_{drift}), and the deviation F_{hom}–F_{drift} in generation 20 for different types of diversity measures for Panels M and N.

Scheme[1]	GBLUP loci (Panel M)			Neutral loci (Panel N)		
	ΔF_{HOM}[2]	ΔF_{drift}[2]	F_{hom}–F_{drift}[3]	ΔF_{HOM}[2]	ΔF_{drift}[2]	F_{hom}–F_{drift}[3]
Drift measures						
G_{VR2}(M,M)	0.0146	0.005	0.147	0.0103	0.0068	0.054
G_{VR2}(M,D)	0.01	0.0069	0.048	0.0101	0.0068	0.05
G_{VR2}(~,M)	0.0109	0.005	0.093	0.0085	0.0059	0.041
G_{VR1}(M,M)	0.0096	0.0056	0.063	0.008	0.0069	0.021
$G_{i(p)}$(M,M)	0.0051	0.0071	−0.053	0.0065	0.0077	−0.031
Homozygosity measures						
$G_{0.5}$(M,M)	0.0008	0.0213	−0.348	0.0073	0.0176	−0.17
G_{ROH}(M,M)	0.0042	0.0091	−0.102	0.0054	0.0088	−0.07
IBD measures						
G_{LA}(M,M)	0.0044	0.0049	−0.009	0.0043	0.0049	−0.01
A(M,~)	0.0072	0.0083	−0.016	0.007	0.0084	−0.021

The target rate of inbreeding for the management of genetic variation was 0.005, and results weigh loci equally irrespective of initial frequency. [1]*See* ***Table 1*** *for scheme names.* [2]*Standard errors <2.5 × 10^{-5}.* [3]*Standard errors <2.2 × 10^{-4}.*

The deviations of F_{hom}–F_{drift} from 0 were significant for all the schemes, for both the SNP-BLUP Panel M and the neutral Panel N, and would imply significant deviations from the classical Eq. (2). The deviation F_{hom}–F_{drift} for $\mathbf{G}_{LA}$(M,M) was closest to the classical expectation, and was closer still after accounting for the degree of non-random mating that was present. Among the remaining schemes **A**(M,~) most closely aligns to classical expectations. The results based on ROH which attempts to mimic IBD appears more similar to $\mathbf{G}_{0.5}$(M,M) which manages homozygosity, where F_{drift} exceeds F_{hom}, although the deviations of the $\mathbf{G}_{0.5}$(M,M) scheme are much larger, with $F_{hom} - F_{drift} = -0.347$ for Panel M which is more than a third of the maximum inbreeding coefficient of 1.

$\mathbf{G}_{VR2}$(M,M), i.e., a commonly used GOC scheme, showed a large deviation opposite to that for $\mathbf{G}_{0.5}$(M,M) with $F_{hom} - F_{drift} = 0.147$ for Panel M, and 0.053 for Panel N, an excess of loss of heterozygosity relative to drift. **Supplementary Information 1** shows this discrepancy must arise due to a covariance between the direction of allele frequency change and initial frequency, with a stronger drift to extremes than would be expected in classical theory. **Figure 3** illustrates this covariance for a randomly chosen replicate, and shows the regression line ($P < 0.001$); for this replicate the difference $F_{hom} - F_{drift} = 0.055$ in Panel N, which arose from a correlation of only 0.040. For $\mathbf{G}_{VR1}$(M,M), which compared to $\mathbf{G}_{VR2}$(M,M) weights the Panel M loci proportional to $2p_{0,k}(1 - p_{0,k})$, this covariance was weaker but was still observed. The result for $\mathbf{G}_{VR2}$(M,D) showed that if the panel used for managing diversity (D) is distinct from that used for SNP-BLUP (M), the covariance in Panel M became similar to that for Panel N, as it is no longer directly managed for its diversity, and the outcome for the unmanaged neutral Panel N was almost identical to $\mathbf{G}_{VR2}$(M,M). The hypothesis that the covariance arises solely as a property of the management by $\mathbf{G}_{VR2}$, rather than as a consequence of the directional selection, was confirmed by the results for $\mathbf{G}_{VR2}$(~,M) where F_{hom} still exceeded F_{drift}. Managing the intensity in scheme $\mathbf{G}_{i(p)}$(M,M) did not remove the covariance but, in contrast to the other "drift" schemes, reversed its sign so that F_{drift} exceeded F_{hom}, which is in accord with the hypothesis that it introduces an increased "cost" of moving toward the extremes compared to $\mathbf{G}_{VR2}$(M,M).

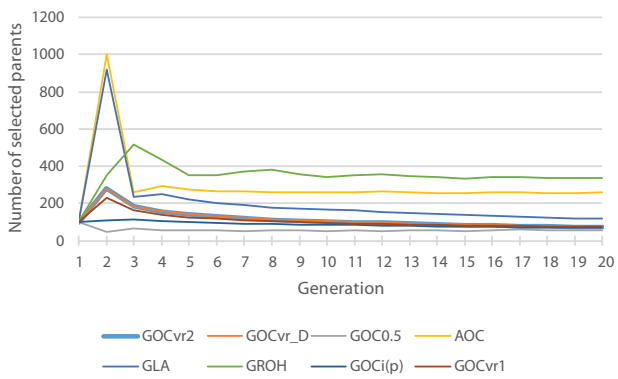


FIGURE 2 | The total number of selected parents for each generation for different breeding schemes. The total is the number of animals with optimal contributions >0 required to achieve a fractional increase in the OC constraint of 0.005.

Managing the Rates of Inbreeding

Table 2 shows ΔF_{drift} and ΔF_{hom} for the different schemes for Panels M and N, and **Figure 4** shows F_{drift} and F_{hom} over time. **Figure 4** shows that $\log(1\text{-}F_{drift})$ is approximately linear with generation for all schemes, in contrast to $\log(1\text{-}F_{hom})$ where some schemes, e.g., $\mathbf{G}_{ROH}$(M,M) show marked curvilinearity.

For $\mathbf{G}_{VR2}$(M,M), ΔF_{drift} for Panel M was directly controlled and was on target at 0.005, but ΔF_{hom} was more than double this target, due to the covariance described above. For Panel N, ΔF_{drift} was greater and ΔF_{hom} was less than observed for Panel M, so the difference was less extreme. The increase in ΔF_{drift}

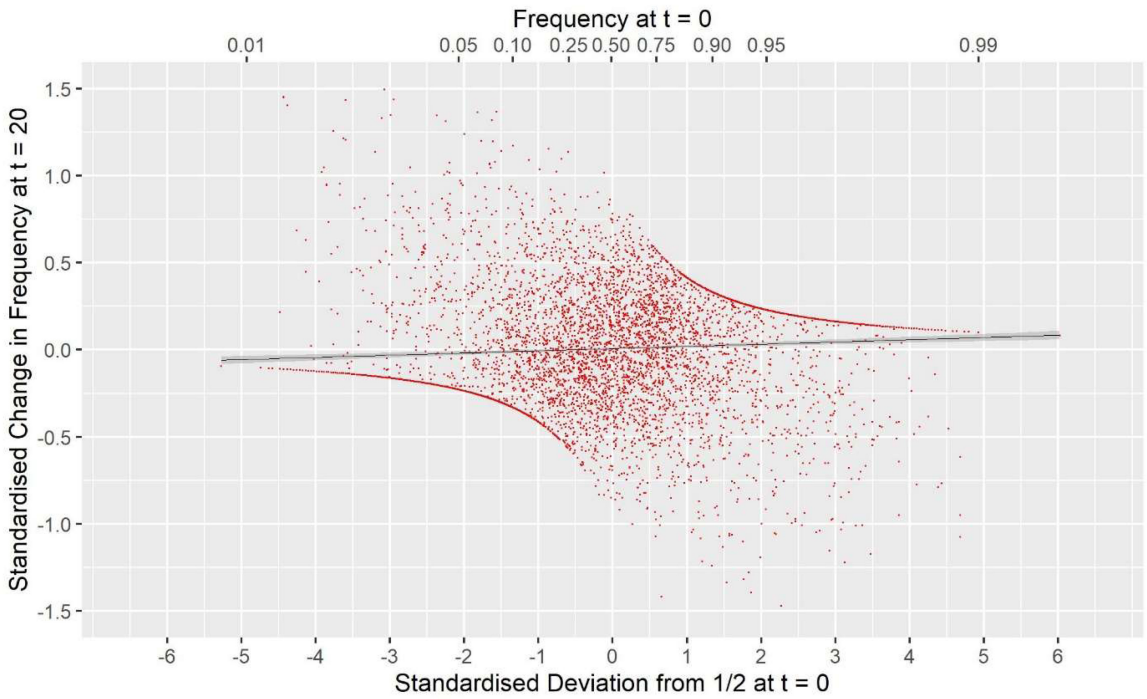


FIGURE 3 | The covariance between the standardized change in allele frequency at t = 20 and the standardize frequency at t = 0 for the 7000 SNP loci in Panel N for a randomly chosen replicate. Standardization is by $\sqrt{p_{0,k}\ (1-p_{0,k})}$ for locus k. The solid black line is the fitted linear regression $y = 0.0083 + 0.0070\times$, with SES 0.0042 and 0.0021, respectively, and a Pearson correlation $r = 0.040$. For this replicate Fdrift = 0.123, Fhom = 0.178, and twice the covariance was 0.0555. The upper x-axis shows the untransformed frequency.

was due to Panel N's LD with QTL that was not accounted for by its LD with Panel M, while the decrease in $\Delta F_{\text{hom}}$was due to the allele frequencies for loci in Panel N being subject to weaker regulation due to their imperfect LD with those in Panel M. The same pattern of differences between ΔF_{drift} and $\Delta F_{\text{hom}}$was observed in a less extreme form with $\mathbf{G}_{\text{VR2}}(\sim,\text{M})$ as here the imperfect LD between Panels M and N is still important but the more favored marker alleles in Panel M change randomly from generation to generation. The outcome for ΔF_{drift} shown in **Table 2** for $\mathbf{G}_{VR1}$(M,M) for Panel M is greater than the target, as F_{drift} and F_{hom} weight all loci in a panel equally, whereas the management weights the drift by $2p_{0,k}(1-p_{0,k})$, consequently the LD with QTL is more weakly constrained for loci with low MAF in Panel M, which is where the impact of the covariance is greatest (**Figure 3**). This also explains the lower ΔF_{hom} observed for $\mathbf{G}_{VR1}$(M,M). The results for $\mathbf{G}_{i(p)}$(M,M) shown in **Table 2** reflect the changed sign in the covariance in that ΔF_{hom} was less than ΔF_{drift}. Unlike $\mathbf{G}_{\text{VR2}}$(M,M), the constraint applied was only indirectly related to F_{drift} or F_{hom} and so the achieved rates were not expected to meet the target, although ΔF_{hom} was close to the target for Panel M.

As with $\mathbf{G}_{i(p)}$(M,M) the simulated management for the measures based on homozygosity, $\mathbf{G}_{0.5}$(M,M) and $\mathbf{G}_{\text{ROH}}$(M,M), did not explicitly control F_{drift} or F_{hom}, However, ΔF_{hom} was close to the desired target for $\mathbf{G}_{\text{ROH}}$(M,M) when measured in both Panels M and N. $\mathbf{G}_{\text{ROH}}$(M,M) showed a curvilinear time trend for F_{hom} mainly due to a negative ΔF_{hom} during the first few generations, after which it increased with time and was rising faster than $\mathbf{G}_{\text{LA}}$(M,M) at the end of the period; in contrast ΔF_{drift} was approximately linear. The accelerating ΔF_{hom} maybe caused by ROHs failing to accumulate inbreeding as haplotypes recombine, so reducing the length of IBD segments below the thresholds implicit in ROH methods, while this older inbreeding is captured by F_{hom}. To test this, the minimum length of a contributing ROH was halved to ∼3.5 from ∼7 Mb but results were nearly identical to those shown in **Table 3** (result not shown). $\mathbf{G}_{0.5}$(M,M) has the highest F_{drift}, because it explicitly promotes allele frequency changes to intermediate frequencies for all loci.

In contrast to all other schemes, ΔF_{drift} for $\mathbf{G}_{\text{LA}}$(M,M) was within 2% of the target for both Panels M and N (see **Table 2**) but was below target for ΔF_{hom} for both panels. The discrepancy for ΔF_{hom} is complicated by the dynamic pattern of the number of parents selected in this scheme (see **Figure 2**), which results in the expected heterozygosity being close to that for random mating in early generations, but ∼0.005 less than random mating in later generation as a result of the degree of non-random mating introduced by the smaller number of parents. Therefore estimating ΔF_{hom} from observed heterozygosity will underestimate the true value and explains a substantial part of the observed deviation from the target value of 0.005. **Figure 4** shows $\mathbf{G}_{\text{LA}}$(M,M) was lowest for F_{drift} and F_{hom} in generation 20 with near constant rates. The results from AOC were qualitatively similar except that both ΔF_{hom} and ΔF_{drift} exceeded the target

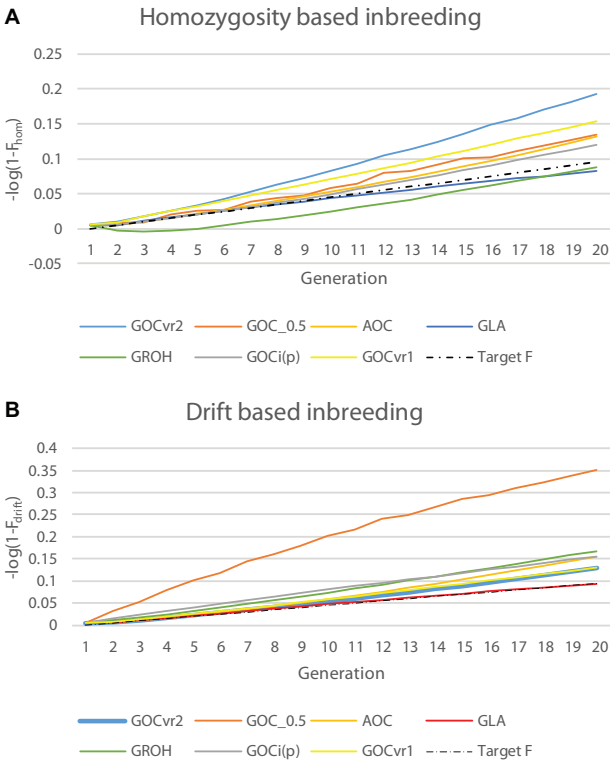


FIGURE 4 | Changes in inbreeding coefficients *F*drift and *F*hom for the neutral loci of Panel N over time plotted on a logarithmic scale where a constant rate of inbreeding results in a linear increase of over time: **(A)** natural logarithm of $(1–F_{hom})$; and **(B)** natural logarithm of $(1–F_{drift})$.

rates by 40% in both panels. This is due to the hitch-hiking of neutral loci with the changes in QTL frequencies arising from the LD generated within families and is unaccounted by using expectations of IBD based on pedigree.

TABLE 3 | Genetic gain (and its SE) after 20 generations of selection expressed in initial genetic standard deviation units, and inbreeding measured by homozygosity for Panel N of neutral loci at generation 20 for comparison.

Scheme	Gain	SE	F_{hom} [1]	F_{drift} [1]
Drift measures				
G_{VR2}(M,M)	7.124	0.002	0.18	0.12
G_{VR2}(M,D)	7.107	0.003	0.17	0.12
G_{VR1}(M,M)	6.680	0.002	0.15	0.12
$G_{i(p)}$(M,M)	7.111	0.003	0.11	0.14
Homozygosity measures				
$G_{0.5}$(M,M)	6.734	0.004	0.13	0.30
G_{ROH}(M,M)	9.099	0.003	0.08	0.15
IBD measures				
G_{LA}(M,M)	7.188	0.002	0.08	0.09
A(M,∼)	9.890	0.003	0.12	0.14

Scheme G_{VR2}*(_,M) is not shown as it was random selection.* [1]*Standard errors* $<3 \times 10^{-3}$.

Genetic Gain

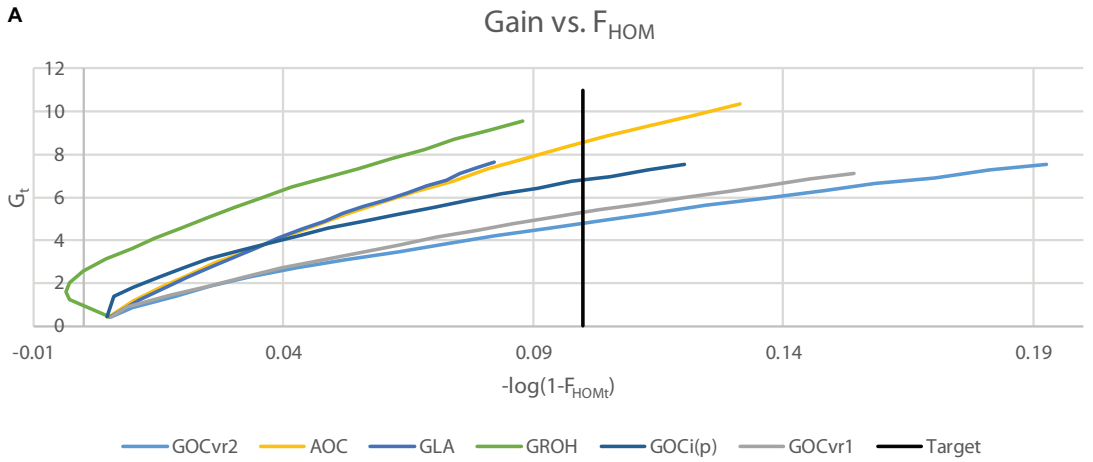
Table 3 shows the genetic gains of the schemes achieved after 20 generations of selection and **Figure 5** shows the gain achieved over time as a function of F_{drift} and F_{hom} for the neutral markers in Panel N. **Figure 5** allows comparisons to be made at the same F_{drift} or F_{hom} and offsets, in part, the unequal rates of inbreeding observed among the different schemes.

The genetic gains were very similar (within 0.3%) for the schemes $\mathbf{G}_{VR2}$(M,M) and $\mathbf{G}_{VR2}$(M,D) where the latter differs only in using a second marker panel for inbreeding management which was unambiguously neutral. Given the small difference in their inbreeding rate at the neutral loci in Panel N (**Tables 2**, **3**), this indicates that separate panels of markers for gain and for diversity is unnecessary for such schemes. The $\mathbf{G}_{LA}$(M,M) scheme yielded significantly more genetic gain than $\mathbf{G}_{VR2}$(M,M), at lower F_{drift} and F_{hom}. $\mathbf{G}_{ROH}$(M,M) and **A**(M,∼) yielded substantially more gain, but their F_{drift} was also higher. The **A**(M,∼) scheme yielded the highest genetic gain of all the schemes compared, but, compared to its closest competitors, $\mathbf{G}_{LA}$(M,M) and $\mathbf{G}_{ROH}$(M,M), it also yielded more F_{drift} and/or F_{hom}.

It is clear from **Figure 5** that the ranking of the schemes for achieved gain differs according to whether drift or homozygosity is considered: e.g., $\mathbf{G}_{ROH}$(M,M) and $\mathbf{G}_{i(p)}$(M,M) schemes yielded relatively high gains given F_{hom}, but relatively low gains given F_{drift}, whereas $\mathbf{G}_{VR2}$(M,M) schemes yielded opposite results with low gains for F_{hom} and relatively high for F_{drift}. The gain for the $\mathbf{G}_{ROH}$(M,M) scheme in early generations was accompanied by negative F_{hom} (**Figure 5A**). $\mathbf{G}_{LA}$(M,M) and **A**(M,∼) schemes performed relatively well as shown in both plots of **Figure 5**, with $\mathbf{G}_{LA}$(M,M) schemes seeming to yield in both plots slightly more gain per unit of inbreeding than **A**(M,∼). Although, the **A**(M,∼) gain is high relative to its inbreeding, the inbreeding rates were substantially larger than the target rate (which can be seen from **Figure 5** by the curves extending far beyond the target). The $\mathbf{G}_{LA}$(M,M) scheme achieves the target rate of inbreeding closely for ΔF_{hom} and ΔF_{drift} (**Table 2**), and simultaneously converts inbreeding efficiently into genetic gain. Moreover, when testing genetic gains in generation 20 of the $\mathbf{G}_{LA}$(M,M) schemes to interpolated gains at the same overall inbreeding (average of F_{hom} and F_{drift}) of the **A**(M,∼) and $\mathbf{G}_{ROH}$(M,M) schemes, the $\mathbf{G}_{LA}$(M,M) scheme yielded the highest gain in 65, respectively, 62 out of 100 replicates; i.e., generation 20 gains of $\mathbf{G}_{LA}$(M,M) were significantly higher than those of **A**(M,∼) and $\mathbf{G}_{ROH}$(M,M) ($P < 0.01$) at the same averaged inbreeding level.

Number of Parents

Figure 2 shows the number of selected parents across the generations and shows that the schemes that use IBD based relationship matrices (**A**, $\mathbf{G}_{LA}$) and $\mathbf{G}_{ROH}$ select most parents. The selected number of parents for $\mathbf{G}_{ROH}$(M,M) may be artificially large due to the additions to the leading diagonal of $\mathbf{G}_{ROH}$ (on average 8.7) to make it positive definite. This process made the $\mathbf{G}_{ROH}$ matrix diagonally dominant, and so reducing $\mathbf{c}'\mathbf{G}_{ROH}\mathbf{c}$ is driven by selecting more parents in order to reduce the impact of these diagonal elements and not about avoiding the



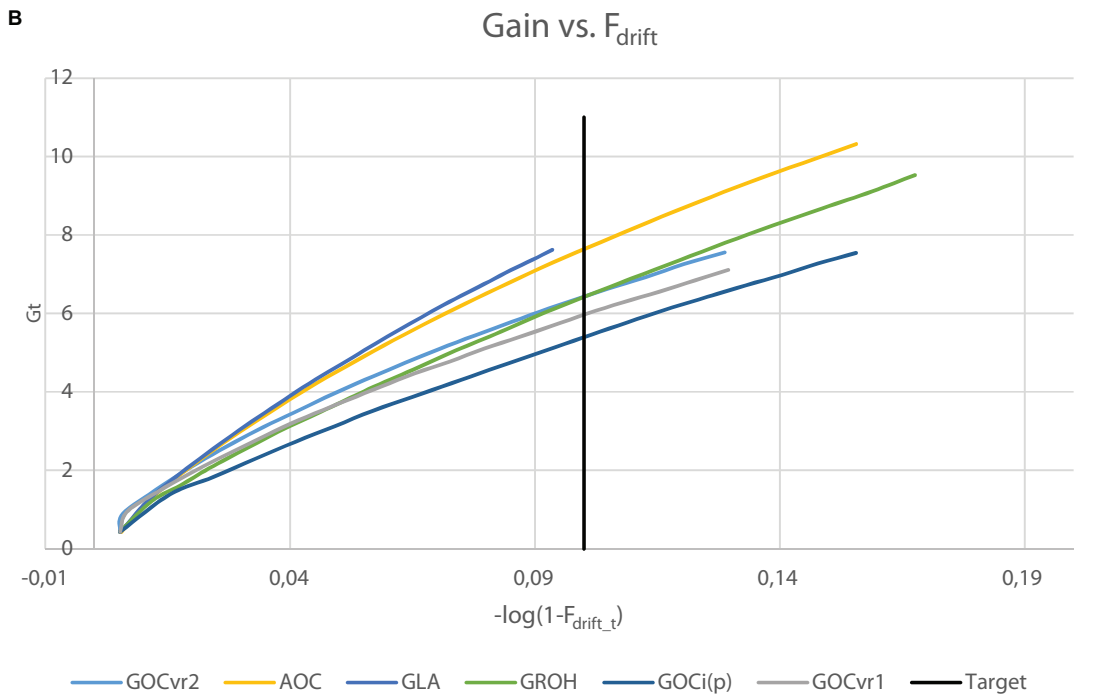


FIGURE 5 | Genetic gain, *Gt* plotted against inbreeding for generations 1–20, where inbreeding is transformed to a logarithmic scale by –log(1-F_t) for F_{hom} **(A)** or F_{drift} **(B)**. For $\Delta F = 0.005$, the target after 20 generations is shown (–log(1-F_t) = 0.1).

selection of related animals. Non-positive definite $\mathbf{G}_{ROH}$ matrices could be inverted to obtain optimal solutions **c**, but these yielded much too high rates of inbreeding (result not shown) probably because optimal contributions **c** were found that resulted in negative $\mathbf{c}'\mathbf{G}_{ROH}\mathbf{c}$, which does not make sense and inbreeding was high and positive. Schemes using matrices constructed by the methodology of VanRaden (2008) ($\mathbf{G}_{VR1}$, $\mathbf{G}_{VR2}$, $\mathbf{G}_{i(p)}$, and $\mathbf{G}_{0.5}$) select fewest parents, implying that they are able to select relatively less related parents by their respective measure, and differences in relationships are relatively large in their respective matrices. Comparing results from **Table 2** and **Figure 2** suggests that the selection of relatively few parents is achieved by making use of the opportunities to induce covariances between allele-frequency-changes and initial frequencies that these schemes offer, which in turn affect the frequencies of heterozygotes.

Genetic Variance

Figure 6 shows the genetic variance for the trait calculated from the true breeding values of the individuals. The $\mathbf{G}_{0.5}$(M,M) scheme loses substantial genetic variance at an early stage, and this relatively low genetic variance is maintained throughout the 20 generations of selection. Therefore striving for allele

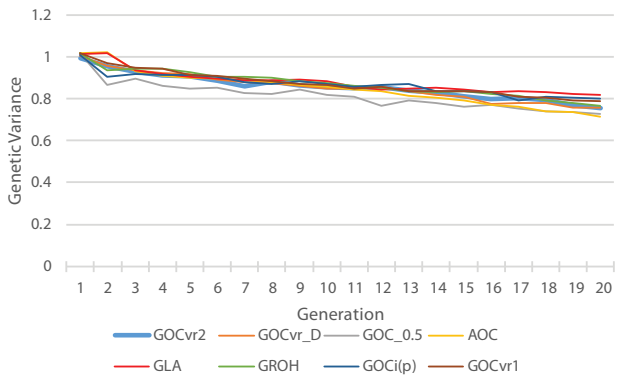


FIGURE 6 | The trait genetic variance of the individuals plotted over time.

frequencies of 0.5 at the loci in Panel M does not maintain variation at the QTL in Panel Q, which is in accord with the results for Panel N in **Table 2**. The relatively low variance for **A**(M,∼) at generation 20 is a consequence of it relatively high genetic gain combined with its relative high rates of inbreeding. By generation 20, the $\mathbf{G}_{\mathrm{LA}}$(M,M) scheme has lost least genetic variance, due to its rates of inbreeding not exceeding the target, and may explain why the $\mathbf{G}_{\mathrm{LA}}$(M,M) scheme is very efficient in turning inbreeding into gain at the end of the selection period (**Figure 5**).

DISCUSSION

Equivalence of Measures F_{hom} and F_{drift}

In the classical work of Wright (1922) two natural measures of inbreeding were introduced concerned with the extent of drift on the one hand (here represented by F_{drift} and $\Delta F_{\mathrm{drift}}$) and heterozygosity on the other (here represented by F_{hom} and ΔF_{hom}), and in classical theory with neutral loci unlinked to QTL these perspectives were identical and directly linked to the occurrence of IBD. The results of this study show that these measures of inbreeding can differ substantially in genomic optimum contribution schemes even when there are no QTL in the genome [$\mathbf{G}_{\mathrm{VR2}}$(∼,M); **Table 2**]. This is because the management in these schemes is commonly directed at the observed homozygosity or drift of the marker loci being monitored. For example, schemes that limit the rate of increase of homozygosity (as represented here by $\mathbf{G}_{0.5}$) induce a negative covariance between the change in allele frequency and the initial frequency, as an excess of minor alleles compared to classical expectations move toward intermediate levels. Conversely schemes managing drift and limiting changes in allele frequency (e.g., using G_{VR2}) induce a positive covariance between change in allele frequency and the initial frequency, as an excess of minor alleles tend to move toward the nearest extreme. Consequently, systematic discrepancies occur between $\Delta F_{\mathrm{drift}}$ and ΔF_{hom}. These discrepancies are a property of the inbreeding management and not of selection *per se*, as they were unaffected by whether random GEBVs were used in the scheme or separate panels of SNPs were used for generating GEBV and management of inbreeding. In contrast to the management using the IBS allele frequencies of monitored markers, when IBD was used either via genomics information (G_{LA}) or approximately (**A**, uninfluenced by markers) the equivalence of $\Delta F_{\mathrm{drift}}$ and ΔF_{hom} was re-established in the simulations, although not with G_{ROH} which is targeted toward IBD but is based on the homozygosity of haplotypes.

The origin of these covariances between allele frequency changes and initial frequencies can be seen when considering the form of the relationship matrix and is explored in detail in **Supplementary Information 1**. The negative covariance arising from $G_{0.5}$ explicitly measures allele frequencies as deviations from 0.5, not from the base frequency $p_{0,k}$ and consequently gains in this measure of diversity (but not necessarily IBD, as discussed later) are obtained by moving frequencies toward 0.5 offsetting any opposing changes prompted by selection objectives. The positive covariance, for example with G_{VR2}, arises because drift of an allele to the more distant extreme is more heavily penalized compared to completely random drift as the GOC with G_{VR2} is constraining the square of the change. This will inevitably promote shifts to the nearest extreme, and more strongly so as p_0 deviates more from ½. Since G_{VR1} is a re-weighting of the loci in G_{VR2} by $w_k/\Sigma_{loci\,k} w_k$ for locus k, where $w_k = 2p_{0,k}(1-p_{0,k})$, placing more weight on frequency changes for loci initially closer to ½, it would be expected the discrepancy between F_{drift} and F_{hom} would be less for G_{VR1} than G_{VR2} as observed in the simulations (see **Table 2** and **Figure 4**). Moving to management using the total intensity applied over time $\left(G_{i(p)}\right)$ penalizes deviations that move toward the extremes more heavily than those toward intermediate frequencies (as $di/dp = [p(1-p)]^{-1/2}$; Liu and Woolliams, 2010), and this changed the sign of the discrepancy although its magnitude was decreased compared to G_{VR2}.

G_{VR2}, which was used by Sonesson et al. (2012), controlled $\Delta F_{\mathrm{drift}}$ and met the target for the panel used (see **Table 2**) but ΔF_{hom} was much greater due to the covariance discussed above. This agreed with the findings of de Beukelaer et al. (2017), where it was suggested that the covariance between change in frequency and its initial value could be the cause of this. However, these authors also reset the allele frequencies for the reference population in the $\mathbf{G}_{VR1}$ matrix every generation to the current generation frequencies, which implies that changes in allele frequency in each generation are constrained without reference to their accumulated change over earlier generations. In a continuous selection scheme, the allele frequency changes of successive generations are positively correlated; thus, although the variance of the change in allele frequency within a generation may have been on target, the variance of the cumulative allele frequency change over generations will exceed the target value due to these positive correlations, as observed in their study. This distinction in methodology will have affected all findings on GOC in the study of de Beukelaer et al. (2017).

Sonesson et al. (2012) found that G_{VR2} schemes achieved their target rate of inbreeding based on IBD using loci with 2N alleles scattered across the genome. Details of the founder populations used in their study were presented in

Sonesson and Meuwissen (2009), which revealed that their SNP-BLUP marker panel was selected for intermediate frequencies in order to mimic a typical SNP-chip marker panel. This is very different from the SNP-BLUP panel used here which was a random sample of whole genome sequence data, and hence dominated by extreme allele frequencies (**Figure 1**). The strength of the covariance underlying the discrepancy between F_{drift} and F_{hom} depends on the distribution of $(p_0 - \frac{1}{2})$, and so in Sonesson et al. (2012) any discrepancy would have been much reduced. In the context of the current results, it was most similar to using G_{VR1} where the intermediate loci are more heavily weighted. Conclusions from these considerations are (i) that the discrepancies between the different measures of rates of inbreeding are extreme in WGS data, due to their extreme allele frequencies (**Figure 1**); and (ii) the discrepancies are a property of the panel used to manage diversity and not the remaining loci, as the IBD-alleles used by Sonesson et al. (2012) have low MAF by construction. Hence, for typical SNPs from chips, the discrepancies between F_{drift} and F_{hom} are expected to be present but smaller than those in **Table 2**.

Management of Diversity

An important aspect of a tool to manage diversity is that it is predictable in meeting its targets, and this can be examined for the marker panel, for the unmanaged neutral markers, and for F_{drift} and F_{hom}. In this respect, G_{VRn} meets the target but only for F_{drift} and only in the marker panel (i.e., not in the unmanaged panel) whereas G_{LA} meets the target (with only minor deviations) for both F_{drift} and F_{hom} for both panels. All others failed to meet the target rate to a greater or lesser degree and would need to be calibrated, possibly in every generation, to meet the targets set at neutral loci. In practice, this would require as realistic as possible simulations of the practical breeding scheme using the current situation as a starting point.

A key management objective in breeding schemes is the efficient generation of gain from the genetic variance in the objectives, and conserving the variation at the (currently) neutral loci, and here the IBD-related schemes were best when compared to F_{drift} or F_{hom} of neutral loci. On an average of F_{drift} and F_{hom}, $\mathbf{G}_{\text{LA}}$ was more efficient than $\mathbf{G}_{\text{ROH}}$, which gave different rates for ΔF_{hom} and ΔF_{drift}, would require regular calibration, and (in the current implementation following de Cara et al., 2013) always required very large number of parents, which in practice would usually demand additional scheme resources. Henryon et al. (2019) observed that using **A** appeared to be more efficient than using G_{VR2}, and this was confirmed here. The differences between schemes using G_{LA} and **A** were small when plotted against F_{drift} or F_{hom} but the G_{LA} scheme was the only scheme tested here that combined high efficiency with rates of inbreeding close to and not exceeding the target rate of inbreeding of 0.005. This supports the conclusion of Sonesson et al. (2012) that genomic selection requires genomic control.

One consequence of entering the genomics era is that the meaning of diversity and its management in practice is more open to discussion, as the pedigree is no longer the only tool to measure and manage it. For example, the number of polymorphic loci could be used as a measure, which might underpin major concerns over the disappearance of known rare alleles in the scheme. Further, in the pedigree inbreeding framework, the measure used is the fraction of variance that is expected to have been lost from the reference base. In the genomic era, if the measure is simply defined as the genetic variance defined by IBS and maximized, there is scope for increasing diversity by the directional selection of loci toward intermediate frequencies as an objective. These measures have been explored elsewhere (see Howard et al., 2017 for a review). In general, attaching values (e.g., selection index weights) to genetic diversity is a very difficult task (e.g., Brisbane and Gibson, 1994; Wray and Goddard, 1994; Goddard, 2009; Jannink, 2010; Howard et al., 2017), which becomes especially clear in view of the aforementioned goals of diversity management, where diversity is required at many (hypothetical) traits simultaneously. Breeders have generally more of an idea about their target rate of inbreeding than on what weight to give to a diversity measure. Although the actual choice of the target rate of inbreeding remains somewhat arbitrary, guidelines have been developed over the years (Woolliams et al., 2015, for a review).

Here, it is argued that an over-riding objective for many populations such as livestock or zoo populations, beyond the breeding goals that underlie the selection on the EBV, is to manage over time the risks associated with the unmeasured attributes of a reference population (e.g., unrecognized deleterious recessives, drift in desirable holistic qualities, epistatic variance). In this respect, all approaches used in this study refer back directly to the established reference (base) population. As mentioned above, other perspectives may be advanced such as increasing the genetic variance at neutral loci by increasing heterozygosity (e.g., de Beukelaer et al., 2017). This could be achieved by the promotion of allele frequency changes toward intermediate values, as exemplified by $G_{0.5}$ in this study, however, this raises issues that require further consideration. Firstly, changes in allele frequency result from multiple copies of a subset of base generation alleles, so increasing frequency is promoting IBD based inbreeding (it is analogous to changing QTL frequency). Secondly, if carried out with a marker panel, then increasing heterozygosity of the marker loci does not necessarily increase heterozygosity among unmonitored neutral loci, which is the objective. In these simulations, the near avoidance of overall loss of heterozygosity in the marker panel by $\text{GOC}_{0.5}$ during selection was accompanied by much greater drift and more loss of heterozygosity in the unmonitored neutral loci than was achieved using IBD based inbreeding management. In contrast, the use of IBD in G_{LA} has information on the unobserved heterozygosity and drift across all the unmonitored genome positions. It remains only a hypothesis that the management of heterozygosity and drift using IBS might perform better than IBD when WGS sequence data is available, with or without selection, although some studies have considered its use (Eynard et al., 2015, 2016; Gómez-Romano et al., 2016). The question how to weigh F_{hom} and F_{drift} across all loci in the genome when a key objective is to manage unknown or unmonitored risks remains open.

While this study has focused on schemes where loss of genetic diversity is managed next to the maximization of genetic

gain, other schemes may be pure conservation schemes, where no genetic change (gain) is desired, but the goals for genetic management are the same; i.e., conserve genetic variation, avoid inbreeding depression, avoid the occurrence of recessive diseases, and avoid random changes in phenotypic traits related to drift from a valued reference population. Strictly, with pure random selection, drift and homozygosity based inbreeding are expected to be the same [Eq. (2); and Falconer and Mackay, 1996]. However, minimisation of allele frequency changes or minimisation of loss of heterozygosity based on using IBS may still result in discrepancies between drift and homozygosity based inbreeding measures arising from the covariances described above. In fact, the potential covariance between the change in allele frequency and the initial frequency is expected to increase, since the inbreeding management term is more important in pure conservation schemes. This would also hold for GOC schemes with selection that aim for an N_e higher than our goal of $N_e = 100$. The greater potential for discrepancy argues for the use of IBD-based measures of relationship (G_{LA}, or a more conservative use of **A**) to maintain diversity in such genetic conservation schemes.

The approach adopted here has not favored genetic variation at some neutral loci more than others *a priori*. Of course, a weighted genomic relationship matrix could be implemented and/or the multiple relationship matrices and associated constraints could be used to simultaneously control the genomic variation in different types of loci (Dagnachew and Meuwissen, 2016; Gómez-Romano et al., 2016). For example, a general **G** matrix covering the entire genome, and an additional **G** matrix controlling genetic diversity at e.g., the major histocompatibility complex, which is essential to the immune response of the animals. Alternatively, regions of the genome may be sought where average heterozygosity is to be increased (reduced) under the assumption that diversity is especially (or not) important in these regions. Regions with known recessive defects may be prioritized for diversity management, but direct inclusion of the known defects in the breeding goal seems more effective in controlling their frequencies. In practice, such regions with special emphasis for diversity management would need to be known *a priori*, and may only be effective if WGS was used for the relationships because, as shown here, what happens in a sample of loci does not necessarily predict what happens at loci outside that subset. Causative alleles of quantitative traits are quite evenly distributed across the genome (Wood et al., 2014), and as argued here the main goals of diversity management address many anonymous, unknown loci and hypothetical traits simultaneously, which makes it very hard to achieve a worthwhile prioritization of genomic regions for diversity management.

CONCLUSION

- Contrary to classic inbreeding theory, inbreeding of unmanaged neutral loci as measured by drift (F_{drift}) and by homozygosity (F_{hom}) can differ very substantially, due to a covariance between the change in allele frequency and its initial frequency, leading to non-zero expected changes in frequency of a sign and magnitude determined by the initial frequency. Discrepancy between F_{drift} and F_{hom} occurs when inbreeding management is based on genomic relationship matrices (or similarity matrices) derived using IBS, but not when derived using IBD, which acts as a unifying concept for F_{drift} and F_{hom}.
- The covariance generated is expected to be larger for WGS data where allele frequencies are extreme with typical MAF close to 0, than for SNP (chip) panels where allele frequencies are generally closer to $^1/_2$.
- The (genomic) selection component of OC schemes does not cause the difference between F_{drift} and F_{hom}.
- Using the same or a different panel for estimating GEBVs than for management of diversity in OC schemes makes only very small differences to genetic gain and the inbreeding in unmonitored neutral loci.
- Measures of genomic relationship can be classified as those based on changes in allele frequency change (e.g., $\mathbf{G}_{\text{VR2}}$) and directed at F_{drift}; those based on homozygosity (e.g., $\mathbf{G}_{0.5}$) and directed at F_{hom}; and IBD based (e.g., $\mathbf{G}_{\text{LA}}$); or combinations of these (e.g., $\mathbf{G}_{\text{ROH}}$). The choice of the relationship matrix depends very much on what objective it should serve.
- OC schemes that limit F_{drift} directly limit allele frequency changes, such as those using $\mathbf{G}_{\text{VR2}}$, result in low ΔF_{drift} at the expense of high ΔF_{hom}. Schemes using $\mathbf{G}_{VR1}$ will be less extreme in this than $\mathbf{G}_{\text{VR2}}$.
- OC schemes that limit ΔF_{hom} (e.g., using $\mathbf{G}_{0.5}$), result in very low ΔF_{hom} at the expense of high ΔF_{drift} but both F_{hom} and F_{drift} may exceed targets at unmonitored neutral loci.
- The OC scheme using $\mathbf{G}_{\text{LA}}$, an IBD based relationship matrix, was the only scheme investigated here that managed homozygosity and drift based inbreeding within the target rate of 0.5%, yielding an effective population size ∼100; for all other schemes, either ΔF_{drift} or ΔF_{hom} or both exceeded their target.
- The OC scheme using $\mathbf{G}_{\text{LA}}$ yielded the highest gain per unit of inbreeding across both measures of inbreeding, closely followed by the scheme using **A**. The latter yielded high gain per unit of F but grossly exceeds target rates of inbreeding.
- The use of $\mathbf{G}_{\text{LA}}$ in practice requires the development of fast algorithms for its calculation.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

TM contributed to study design, performed the simulations, and wrote the draft manuscript. AS developed the simulation software and contributed to discussions and the writing of the manuscript. GG contributed to discussions and the writing of the manuscript. JW contributed to study design, alternative schemes and methods, and discussions

and writing of the manuscript. All authors approved the final version of the manuscript.

FUNDING

We are grateful for funding from the Norwegian Research Council (Grant 226275/E40). JW would like to acknowledge funding from the European Commission under Grant Agreement 677353 (IMAGE) and BBSRC Institute Strategic Programe BBS/E/D/30002275.

ACKNOWLEDGMENTS

We would like to thank three reviewers for their very helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00880/full#supplementary-material

REFERENCES

Brisbane, J. R., and Gibson, J. P. (1994). Balancing selection response and rate of inbreeding by including genetic relationships in selection decisions. *World Congr. Genet. Appl. Livest. Prod.* 19:135.

Charlier, C., Coppieters, W., Rollin, F., Desmecht, D., Agerholm, J. S., Cambisano, N., et al. (2008). Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat. Genet.* 40, 449–454. doi: 10.1038/ng.96

Dagnachew, B. S., and Meuwissen, T. H. E. (2016). A fast iterative algorithm for large scale optimal contribution selection. *Gen. Sel. Evol.* 48:70.

de Beukelaer, H., Badke, Y., Fack, V., and deMeyer, G. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. *Genetics* 206, 1127–1138. doi: 10.1534/genetics.116.194449

de Cara, M. A. R., Villanueva, B., Toro, M. A., and Fernández, J. (2013). Using genomic tools to maintain diversity and fitness in conservation programmes. *Mol. Ecol.* 22, 6091–6099. doi: 10.1111/mec.12560

Eynard, S. E., Windig, J. J., Hiemstra, S. J., and Calus, M. P. (2016). Whole-genome sequence data uncover loss of genetic diversity due to selection. *Genet. Sel. Evol.* 48:33.

Eynard, S. E., Windig, J. J., Leroy, G., van Binsbergen, R., and Calus, M. P. (2015). The effect of rare alleles on estimated genomic relationships from whole genome sequence data. *BMC Genet.* 16:24. doi: 10.1186/s12863-015-0185-0

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction To Quantitative Genetics*. Harlow: Pearson Education Limited.

Fernandez, J., Villanueva, B., Pong-Wong, R., and Toro, M. A. (2005). Efficiency of the use of pedigree and molecular marker information in conservation programs. *Genetics* 170, 1313–1321. doi: 10.1534/genetics.104.037325

Fernando, R. L., and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Gen. Sel. Evol.* 21, 467–477.

Goddard, M. E. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Gómez-Romano, F., Villanueva, B., Fernández, J., Woolliams, J. A., and Pong-Wong, R. (2016). The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. *Genet. Sel. Evol.* 48:2.

Henryon, M., Liu, H., Berg, P., Su, G., Nielsen, H. M., Gebregewergis, G. T., et al. (2019). Pedigree relationships to control inbreeding in optimum-contribution selection realise more genetic gain than genomic relationships. *Genet. Sel. Evol.* 51:39.

Holsinger, K. E., and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* 10, 639–650. doi: 10.1038/nrg2611

Howard, J. T., Pryce, J. E., Baes, C., and Maltecca, C. (2017). Invited review: inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. *J. Dairy Sci.* 100, 6009–6024. doi: 10.3168/jds.2017-12787

Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42:35.

Keller, M. C., Visscher, P. M., and Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237–249. doi: 10.1534/genetics.111.130922

Kinghorn, B. P. (1980). The expression of recombination loss in quantitative traits. *J. Anim. Breed. Genet.* 97, 138–143. doi: 10.1111/j.1439-0388.1980.tb00919.x

Legarra, A. (2016). Comparing estimates of genetic variance across different relationship models. *Theor. Popul. Biol.* 107, 26–30. doi: 10.1016/j.tpb.2015.08.005

Leinster, T., and Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology* 93, 477–489. doi: 10.1890/10-2402.1

Li, C. C., and Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* 5, 107–117.

Liu, A. Y., and Woolliams, J. A. (2010). Continuous approximations for optimizing allele trajectories. *Genet. Res.* 92, 157–166. doi: 10.1017/s0016672310000145

Luan, T., Yu, X., Dolezal, M., Bagnato, A., and Meuwissen, T. H. (2014). Genomic prediction based on runs of homozygosity. *Genet. Sel Evol.* 46:64. doi: 10.1016/j.cancergen.2018.04.038

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., et al. (2008). Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 83, 359–372.

Meuwissen, T. H. E. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *J. Anim. Sci.* 75, 934–940.

Meuwissen, T. H. E., and Goddard, M. E. (2010). The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. *Genetics* 185, 1441–1449. doi: 10.1534/genetics.110.113936

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Meuwissen, T. H. E., Luan, T., and Woolliams, J. A. (2011). The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *J. Anim. Breed. Genet.* 128, 429–439. doi: 10.1111/j.1439-0388.2011.00966.x

Pong-Wong, R., and Woolliams, J. A. (2007). Optimisation of contribution of candidate parents to maximise genetic gain and restricting inbreeding using semidefinite programming. *Genet. Sel. Evol.* 39, 3–25.

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.* 11, 800–805. doi: 10.1038/nrg2865

Robertson, A. (1965). The interpretation of genotypic ratios in domestic animal populations. *Anim. Prod.* 7, 319–324. doi: 10.1017/s0003356100025770

Rodríguez-Ramilo, S. T., Fernández, J., Toro, M. A., Hernández, D., and Villanueva, B. (2015). Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population. *PLoS One* 10:e0124157. doi: 10.1371/journal.pone.0124157

Sonesson, A. K., and Meuwissen, T. H. E. (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genet. Sel. Evol.* 41:37.

Sonesson, A. K., Woolliams, J. A. W., and Meuwissen, T. H. E. (2012). Genomic selection requires genomic control of inbreeding. *Genet. Sel. Evol.* 44:27.

Toro, M. A., Silio, L., Rodriganez, J., and Rodriguez, C. (1998). The use of molecular markers in conservation programmes of live animals. *Genet. Sel. Evol.* 30:585. doi: 10.1186/1297-9686-30-6-585

Toro, M. A., Villanueva, B., and Fernandez, J. (2014). Genomics applied to management strategies in conservation programmes. *Livestock Sci.* 166, 48–53. doi: 10.1016/j.livsci.2014.04.020

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Villanueva, B., Pong-Wong, R., Fernandez, J., and Toro, M. A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* 83, 1747–1752. doi: 10.2527/2005.8381747x

Wang, J. (2001). Optimal marker-assisted selection to increase the effective size of small populations. *Genetics* 157, 867–874.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186.

Woolliams, J. A., Berg, P., Dagnachew, B. S., and Meuwissen, T. H. E. (2015). Genetic contributions and their optimization. *J. Anim. Breed. Genet.* 132, 89–99. doi: 10.1111/jbg.12148

Wray, N. R., and Goddard, M. E. (1994). Increasing long term response to selection. *Genet. Sel. Evol.* 26:431. doi: 10.1186/1297-9686-26-5-431

Wright, S. (1922). Coefficients of inbreeding and relationships. *Amer. Nat.* 56, 330–338.

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.

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