frontiers | Research Topics

Insights in evolutionary and population genetics 2022

Edited by
Samuel A. Cushman, Rongling Wu and
Rosane Garcia Collevatti

Published in
Frontiers in Genetics
Frontiers in Ecology and Evolution

ISSN 1664-8714
ISBN 978-2-8325-3013-9
DOI 10.3389/978-2-8325-3013-9

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: frontiersin.org/about/contact

Insights in evolutionary and population genetics: 2022

Topic editors

Samuel A. Cushman — Forest Service, United States Department of Agriculture (USDA), United States
Rongling Wu — The Pennsylvania State University (PSU), United States
Rosane Garcia Collevatti — Universidade Federal de Goiás, Brazil

Citation

Cushman, S. A., Wu, R., Collevatti, R. G., eds. (2023). *Insights in evolutionary and population genetics: 2022*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3013-9

Table of contents

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 15 July 2022
DOI 10.3389/fevo.2022.895056

OPEN ACCESS

EDITED BY
Alison G. Nazareno,
Federal University of Minas Gerais,
Brazil

REVIEWED BY
Alessandro Alves-Pereira,
State University of Campinas, Brazil
Alexandre Magno Sebbenn,
Instituto Florestal, Brazil

*CORRESPONDENCE
Danilo E. Bustamante
danilo.bustamante@untrm.edu.pe

SPECIALTY SECTION
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 12 March 2022
ACCEPTED 29 June 2022
PUBLISHED 15 July 2022

CITATION
Bustamante DE, Motilal LA,
Calderon MS, Mahabir A and Oliva M
(2022) Genetic diversity and
population structure of fine aroma
cacao (*Theobroma cacao* L.) from
north Peru revealed by single
nucleotide polymorphism (SNP)
markers.
Front. Ecol. Evol. 10:895056.
doi: 10.3389/fevo.2022.895056

Genetic diversity and population structure of fine aroma cacao (*Theobroma cacao* L.) from north Peru revealed by single nucleotide polymorphism (SNP) markers

Danilo E.Bustamante [1,2]*, Lambert A. Motilal [3], Martha S. Calderon [1,2], Amrita Mahabir [3] and Manuel Oliva [1]

[1]Instituto de Investigación para el Desarrollo Sustentable de Ceja de Selva (INDES-CES), Universidad Nacional Toribio Rodríguez de Mendoza, Chachapoyas, Peru, [2]Instituto de Investigación en Ingeniería Ambiental (IIIA), Facultad de Ingeniería Civil y Ambiental (FICIAM), Universidad Nacional Toribio Rodríguez de Mendoza, Chachapoyas, Peru, [3]Cocoa Research Centre, The University of the West Indies, St. Augustine, Trinidad and Tobago

Cacao (*Theobroma cacao* L.) is the basis of the lucrative confectionery industry with "fine or flavour" cocoa attracting higher prices due to desired sensory and quality profiles. The Amazonas Region (north Peru) has a designation of origin, Fine Aroma Cacao, based on sensory quality, productivity and morphological descriptors but its genetic structure and ancestry is underexplored. We genotyped 143 Fine Aroma Cacao trees from northern Peru (Bagua, Condorcanqui, Jaén, Mariscal Cáceres, and Utcubamba; mainly Amazonas Region), using 192 single nucleotide polymorphic markers. Identity, group, principal coordinate, phylogenetic and ancestry analyses were conducted. There were nine pairs of matched trees giving 134 unique samples. The only match within 1,838 reference cacao profiles was to a putative CCN 51 by a Condorcanqui sample. The "Peru Uniques" group was closest to Nacional and Amelonado-Nacional genetic clusters based on F_{ST} analysis. The provinces of Bagua and Utcubamba were genetically identical (D_{est} = 0.001; P = 0.285) but differed from Condorcanqui (D_{est} = 0.016–0.026; P = 0.001–0.006). Sixty-five (49%) and 39 (29%) of the Peru Uniques were mixed from three and four genetic clusters, respectively. There was a common and strong Nacional background with 104 individuals having at least 30% Nacional ancestry. The fine aroma of cacao from Northern Peru is likely due to the prevalent Nacional background with some contribution from Criollo. A core set of 53 trees was identified. These findings are used to support the continuance of the fine or flavour industry in Peru.

KEYWORDS

core collection, fine or flavour cocoa, genetic structure, group differentiation, Nacional ancestry, north Peru, phylogeny, Peruvian Amazonas region

Introduction

Domestication and use of *Theobroma cacao* L. (cacao; chocolate tree) dates back to ∼5,000 years from ruins of the Chinchipe culture, Palanda, south-eastern Ecuador and Montegrande, Jaen, Peru (Valdez, 2013; Ochoa, 2017; De la Fuente, 2018; Olivera-Núñez, 2018; Zarrillo et al., 2018). Cacao is used to refer to the plant while cocoa is used for the fermented and dried seeds and their processed products. Cacao is a tropical dicot Malvaceae tree (Alverson et al., 1999; Bayer et al., 1999) native to the Amazon basin of South America (Toxopeus, 1985; Motamayor and Lanaud, 2002; Bartley, 2005). The fruits produce seeds that are used in the pharmaceutical and cosmetic industries but primarily as the raw ingredients for the multibillion dollar chocolate industry (Oddoye et al., 2013; Wickramasuriya and Dunwell, 2018). The consumption of chocolate and its products is estimated to increase by 3% each year (Wickramasuriya and Dunwell, 2018) and acts as the main economic driver of global cacao farming (Tacer-Caba, 2019).

Cacao crops are critical for local economies of about 6 million smallholder farmers in Latin America, Africa, and Asia (Rice and Greenberg, 2000; Beg et al., 2017). Peru produced 160,289 metric tonnes of cocoa in 2020 making it the 9th largest producer of cocoa worldwide (FAO, 2022). In the Amazonas region of Peru, cacao is the second most economically important crop with a cultivated and harvested area of 13,416.83 ha (Instituto Nacional de Estadística e Informática [INEI], 2012). In this region, three provinces are the main producers of cacao: Bagua with the highest cocoa production (75%), followed by the Utcubamba, and Condorcanqui provinces (Torres-Armas and Gonzáles-Castro, 2018). The discovery of high-yielding and disease-resistant varieties is needed to support the growing global cacao industry (Goenaga et al., 2009; Phillips-Mora et al., 2013). The conservation and utilisation of cacao genetic diversity are crucial for the sustainable cultivation of cacao (Zhang and Motilal, 2016; Laliberté et al., 2018).

The cocoa industry recognises "bulk cocoa" and "fine or flavour cocoa" with the latter garnering a higher premium price. While bulk cacao still contributes to more than 80% of worldwide production (Wickramasuriya and Dunwell, 2018), there has been an increase in demand for fine or flavour chocolate, along with consumer appreciation for the traditional histories and origin of native cacao varieties (Mejía et al., 2021). Peru has been designated a 75% producer and exporter of fine flavour cocoa (International Cocoa Organization [ICCO], 2021) and is thus well poised to capitalise on this consumer base. Cacao from the Peruvian Amazonas region currently has a designation of origin, namely Fine Aroma Cacao, based on its peculiar characteristics in terms of its sensory quality (aroma and flavour) (Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual [INDECOPI], 2016). These qualities have given a high value and demand which strongly improve the competitiveness of Peruvian Amazonas cocoa in the foreign market (Oliva and Maicelo-Quintana, 2020). Five groups of cacao (Bagüinos, Cajas, Indes, Toribianos, Utkus) were identified according to these sensory features, in addition to productivity and morphological descriptors (Oliva-Cruz, 2020). Sensory evaluation of Bagua type cacao determined that this upper Amazon variety differed from the native "Chuncho" cacao found in Quillabamba, Cusco (Céspedes-Del Pozo et al., 2018; Mejía et al., 2021). The Indes and Bagüinos morphotypes had the best floral and fruity sensory characteristics and the highest dry weight and number of seeds (Oliva-Cruz, 2020).

Bulk cocoa traditionally comes from Forastero cacao while "fine or flavour" cocoa can be obtained from Criollo, Nacional and some Trinitario varieties (Pridmore et al., 2000). Cacao was traditionally classed as Criollo, Forastero and Trinitario varieties, based on morphological and agronomical traits with the latter variety being thought to be a hybrid of the former two (Toxopeus, 1985; Pridmore et al., 2000). Forastero encompassed a range of cacao types including the Amelonado variety responsible for the basis of the West African bulk cocoa industry and the Nacional variety from Ecuador known for its fine Arriba flavour. The Refractario variety also from Ecuador arose out of a mass field selection program in the 1920s for witches' broom disease resistance (Pound, 1938, 1943; Bartley, 2001).

A variety of molecular approaches have enabled better separation and understanding of the true genetic diversity and varietal classification than the traditional names and industry convention. A review of these molecular approaches can be found in Livingstone et al. (2012), Motilal et al. (2017), and Everaert et al. (2020). Genetic diversity is higher when there are unique samples that increase differentiation within and among groups. Accurate identity analysis is, however, dependent on the number of markers, as well as, the composition of the marker set used for both microsatellite markers (Motilal et al., 2009) and single nucleotide polymorphism (SNP) markers (Mahabir et al., 2020).

The use of microsatellite markers is currently being supplanted by SNP markers especially for large genetic diversity studies. Genotyping of cacao germplasm with SNPs has been performed using the novel integrated fluid circuit (IFC) technology (Osorio-Guarín et al., 2017), which increased the throughput per run, simplified setup of reactions, and decreased the running cost (Xu, 2016). Lately, the analysis of the genetic diversity and population structure of cacao have used a set of reduced and informative SNP markers (Singh and Singh, 2015; Cosme et al., 2016; Osorio-Guarín et al., 2017; Mahabir et al., 2020; Wang et al., 2020). The identification and authentication of fine flavour cacao varieties have also employed SNPs (Fang et al., 2014; Arevalo-Gardini et al., 2019).

Genetic clustering of cacao was clarified by Motamayor et al. (2008) who used 106 microsatellite markers to identify ten genetic groups (Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay, and Purús) in the Amazon basin of South America. The clustering of these groups has been supported and refined by Thomas et al. (2012) and Nieves-Orduña et al. (2021). Five of the 10 genetic clusters (Contamana, Iquitos, Marañon, Nacional, and Nanay) occur in Peru (Motamayor et al., 2008). Nieves-Orduña et al. (2021) identified 23 chloroplast microsatellite haplotypes on a sample of 233 cacao plants with the highest variation being found in western Amazonia; particularly in the north-western Amazon with Peru having seven unique haplotypes. The genetic clustering of cacao is expected to change as more wild natural stands of cacao are explored in the Amazon. North-eastern Peru hosts a wide diversity and genetic variability of cacao that is under-explored (Motamayor et al., 2008; Thomas et al., 2012). Two traditional fine or flavour varieties in Peru are the small-seeded variety known as Chuncho from the Urubamba valley in southern Peru; and the "Piura Porcelana" variety with large pale seeds mainly cultivated in Piura, Amazonas, and Cajamarca provinces of northern Peru (Arevalo-Gardini et al., 2019). Céspedes-Del Pozo et al. (2018), using 96 single nucleotide polymorphism (SNP) markers reported that the native cacao variety "Chuncho" –from La Convencion, Cusco in southern Peru – was distinct but closest to the Contamana population, Beni population (unique population from Beni River in Bolivia, Zhang et al., 2012), and cacao from the Madre de Dios region. "Piura Porcelana" formed an immediate sister clade to the Nacional group (Arevalo-Gardini et al., 2019).

Additionally, Chia-Wong et al. (2018) tried to assess about 80 fine or flavour trees from the five principal cacao regions of Peru (Amazonas, Cusco, San Martin, Piura, and Huánuco) with 18 microsatellites but the amplification was unsuccessful. Zhang et al. (2006a), using 15 microsatellites, demonstrated that cacao in Huallaga and Ucayali Valleys were distinct groups. The Huallaga farmer selections were shown to be mainly hybrids of Trinitario and Upper Amazon Forastero accessions (Zhang et al., 2011).

Saavedra-Arbildo et al. (2018) demonstrated from fruit and seed morphology that the cacao in the Peruvian regions of Amazonas, Cusco and Piura were similar in thickness of fruit wall, fruit length, water content of testa, and seed width but differed in depth of primary furrows, fruit mass, seed mass, fruit width, number of seeds, dry mass of seeds, seed length, and seed thickness. In addition, the northern regions of Amazonas and Piura appeared more similar to each other than Cusco although all three areas were differentiated on the basis of number of seeds and seed length with the Piura region having the greatest proportion of white seeds in fruits that were generally elliptic-obovate with obtuse apices and little to no rugosity (Saavedra-Arbildo et al., 2018).

There is a scarcity of recent work on the phenotypic and genetic diversity of cacao in Peru, far less for northern Peru. In addition, the use of the current SNP marker technology is limited to a few studies. Studies on the genetic diversity of cacao and especially fine aroma cacao in northern Peru are lacking. The goal of this study is to determine the genetic uniqueness, genetic diversity and ancestry of Fine Aroma Cacao from the Peruvian Amazonas region by SNP genotyping. In addition, we examined if three provinces (Bagua, Condorcanqui, and Utcubamba) were genetically distinct and harboured new cacao genetic clusters. The resultant information is expected to be a significant addition to our understanding of the genetic diversity of cacao in Peru and how it can be leveraged to bolster the fine flavour status in Peru.

Materials and methods

Sample collection

A total of 143 trees (15–20 years old) of Fine Aroma Cacao were sampled mainly from farmers' fields in three provinces of the Amazonas region, in northern Peru (Bagua, Condorcanqui, Utcubamba; **Supplementary Table 1** and **Figure 1**) and were deposited in the herbarium of Universidad Nacional Toribio Rodríguez de Mendoza (KUELAP), Peru (Thiers, 2016). A permit for scientific research on wild flora (RDG N° D000319-2020-MINAGRI-SERFOR-DGGSPFFS, with authorisation code N° AUT-IFL-2020-051) was provided by Servicio Nacional Forestal y de Fauna Silvestre (SERFOR). For each site, the date, time, and GPS coordinates were recorded. The 143 test trees from northern Peru were compared to reference profiles of cacao accessions belonging to the 10 genetic clusters of Motamayor et al. (2008) as well as Trinitario and Refractario accessions. A maximum of 1,838 reference accessions from the International Cocoa Genebank Trinidad were used. Reference profiles are maintained and curated by the Cocoa Research Centre (CRC), The University of the West Indies.

Single nucleotide polymorphism genotyping and curation

Tissue samples were taken from the distal regions of healthy cacao leaves and stored in pre-labelled 1.5 mL Safelock Eppendorf tubes containing silica gel desiccant. Six leaf discs (6 mm diameter) were prepared from each test plant using the BioArk leaf collection kit from LGC Biosearch Technologies. The plates were shipped to LGC Genomics, United Kingdom for DNA extraction and SNP genotyping using their proprietary KASP chemistry. Genotyping was performed at 192 SNP sites from flanking sequences provided by the Cocoa Research Centre (CRC) of The University

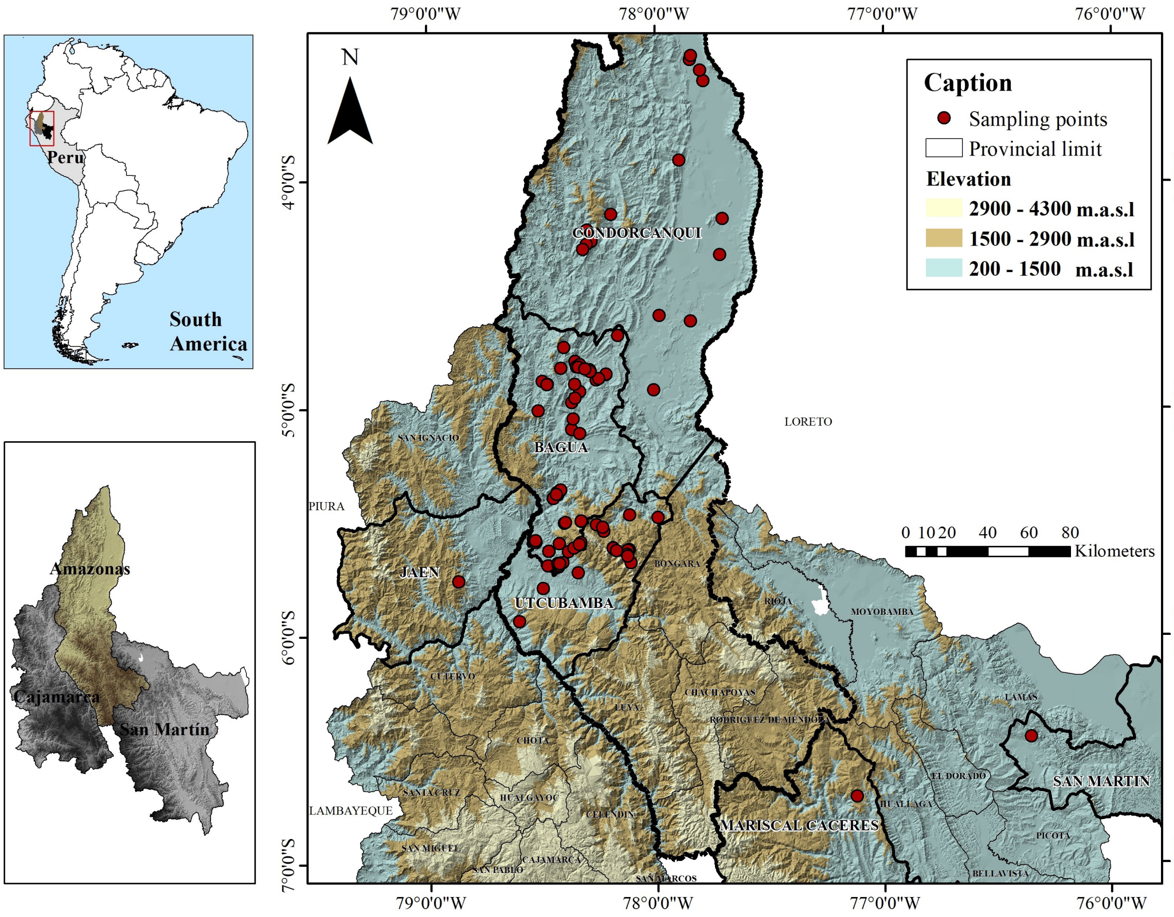


FIGURE 1
Distribution of collected Fine Aroma Cacao samples from northern Peru. The national, provincial and district boundaries were obtained from the Geoportal of the National Geographic Institute of Peru (IGN) in shapefile format with a DATUM WGS 1984 for illustrative purposes only.

of the West Indies (Motilal et al., 2017; Mahabir et al., 2020; **Supplementary Table 2**). Returned multilocus data from LGC was curated by removing SNPs and samples with more than 7% missing data. This is expected to reduce the impact of missing data on the genetic analyses. Seven of the 192 SNPs (TcSNP 0456, 0701, 1,038, 1,156, 1,229, 1,408, 1,457) had 100% missing data and were removed from subsequent analyses.

Software and analysis overview

Multilocus SNP profiles were analysed using GenAlEx v6.502 (Peakall and Smouse, 2006, 2012). This software was used for frequency analysis, group differentiation tests, principal coordinate analysis (PCoA) and to prepare in files for other programs. Identity analyses were conducted in Cervus v3.0 (Kalinowski et al., 2007), phylogenetic analysis in DARwin v6 (Perrier et al., 2003; Perrier and Jacquemoud-Collet, 2006), ancestry analysis in STRUCTURE v2.3.4 (Pritchard et al., 2000) and core collection identification in PowerCore v1.0 (National Institute of Agricultural Biotechnology, 2006). Statistical tests to determine if there were significant differences in genetic parameters were conducted in MedCalc Statistical Software v12.7.7 (MedCalc Software bvba, 2013).

Identity analysis

Identity analyses were conducted in Cervus v3.0 (Kalinowski et al., 2007). A minimum of 170 matching loci with a flexibility mismatch of 5 loci was applied to identify possible groups of matched samples among the data matrix of 185 SNPs/143 Peru test trees (**Supplementary Table 3**). Missing data occurred at 0–7 SNPs with an average of 0.60% (standard error = 0.05) in the entire dataset. Members within a group are equivalent to each other but not equivalent to members of other groups. One member of each group was retained to obtain a maximal list of unique Peru samples

(Peru Uniques). Identity analysis was conducted between the Peru Uniques dataset and 1,838 CRC reference profiles mainly from the International Cocoa Genebank Trinidad. The reference profiles had data at 175 SNPs so identity analysis was conducted using a minimum of 160 matching loci with a flexibility mismatch of 5 loci. In this dataset, missing data was present at 0–42 SNPs per sample (mode = 6) with an average of 7.69% (standard error = 0.16) in the entire dataset. The probability of identity among siblings (PID_{SIB}), was obtained to estimate the chance of a false match. The PID_{SIB} is the probability that two siblings drawn at random from a population have identical genotypes (Evett and Weir, 1998; Waits et al., 2001) and was recommended to be used in cacao (Zhang et al., 2006b).

Frequency analysis

The 134 Peru Uniques had 8 and 26% missing data at TcSNP0230 and TcSNP1350, respectively, over the maximal set of 185 SNPs. In addition, three monomorphic SNPs (TcSNP: 0097, 0383, 1,158) were present. These five SNPs were removed so that missing data was present at 0–14 SNPs per sample (mode = 0) with an average of 0.41% (standard error = 0.05) in the entire dataset. Frequency analysis was conducted in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012) to obtain descriptive genetic measures of number of effective alleles (N_{rme}); Shannon's Information Index (*I*); observed, expected and unbiased expected heterozygosities (H_o, H_e, uH_e, respectively); the fixation index (F); and individual heterozygosities (H_{ind}) of the sampled trees for each sampled provinces and the Peru Uniques.

Principal coordinate analysis

Principal coordinate analysis was conducted on the set of Peru Uniques in relation to 390 reference accessions (40 Amelonado, 8 Contamana, 12 Criollo, 17 Curaray, 56 Guiana, 32 Iquitos, 70 Marañon, 40 Nacional, 60 Nanay, 5 Purús, 25 Amelonado/Nacional hybrids, and 25 Amelonado/Criollo hybrids) using 170 SNPs. Population references are from selected accessions with exclusive membership to their respective genetic clusters. Similarly hybrid references were selected based on contributions from only the two required genetic clusters. Accessions and SNPs were chosen to minimise missing data. In this dataset, missing data was present at 0–14 SNPs per sample (mode = 0) with an average of 0.44% (standard error = 0.03) in the entire dataset. The analysis in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012) implemented a standardised linear genetic distance.

Phylogenetic analysis

Phylogenetic analysis was performed on the same dataset as for the PCoA in DARwin v6 (Perrier et al., 2003; Perrier and Jacquemoud-Collet, 2006). The program accepts allelic data and creates a simple matching dissimilarity index. Missing data was set at 50, 70, or 90% with the default pairwise allele deletion to construct dissimilarity matrices with 1,000 bootstraps. Tree construction employed the weighted Neighbor-Joining algorithm with 1,000 bootstrap replicates. Bootstrap values > = 70% were displayed on the trees.

Group differentiation tests

An analysis of molecular variance (AMOVA) was conducted using the same dataset as for the PCoA using 999 permutations in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012). Group differentiation based on Jost D_{est} statistic (Jost, 2008, 2009) was conducted on a refined dataset. This dataset involved the same reference groups as for the PCoA but with 136 SNPs to get less than 6% missing data per group and with an average of 0.13% (standard error = 0.01) missing data within the dataset. In addition, the Peru Uniques was decomposed into provincial groups that contained at least five samples. The provinces of Bagua ($n = 16$), Condorcanqui ($n = 24$) and Utcubamba ($n = 91$) were retained. If members of duplicate groups were present, only one sample per group province was retained. In this dataset, there was also less than 6% missing data per group and with an average of 0.12% (standard error = 0.01) missing data. The D_{est} pairwise calculations were performed in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012) using 999 permutations and 999 bootstraps. Phylogenetic clusters in the collected samples were identified and Jost D_{est} was used to determine if the clusters were separate groupings. Datasets were examined for private alleles.

Ancestry analysis

Population structure of the 143 samples was determined via the model-based clustering method implemented in STRUCTURE v2.3.4 (Pritchard et al., 2000). Reference accessions that represent the 10 genetic clusters identified by Motamayor et al. (2008) were cloned to obtain a sample size of 200 for each population. An initial run using number of populations (*K*) from nine to 14 [the expected 10 of Motamayor et al. (2008) plus one more than the number of expected clusters in the collected samples] was conducted. A dataset of 154 SNPs was used to obtain minimal missing data. An admixture model with an inferred alpha value, independent allele frequency with 100,000 burnins and 200,000 Markov Chain Monte Carlo (MCMC) repetitions was used with 10 iterations at each *K* value. The optimal *K* was selected based on best differentiation of samples to maintain Motamayor et al. (2008) grouping and on the *ad hoc* method of Evanno et al. (2005). Then with a maximal dataset of 170 SNPs in the Peru samples and cloned population references an admixture model with an inferred alpha value, independent allele frequency with 300,000 burnins and 600,000 MCMC repetitions was used at the chosen *K* with 10 iterations. The run with the most positive ln P(D) was chosen to represent the ancestral background. A minimum level of 5% was used as evidence of the presence of a genetic group. A minimum level of 95% without a 5% level in any other group was used to establish exclusive membership to a genetic group. The distribution of the predominant ancestral group(s) for the Bagua, Condorcanqui and Utcubamba provinces was tested for equivalence using the comparison of proportions test in MedCalc Statistical Software v12.7.7 (MedCalc Software bvba, 2013).

Core collection identification

The Peru Uniques typed at the maximal number of SNPs underwent core selection in PowerCore v1.0 (National Institute of Agricultural Biotechnology, 2006) under its heuristic algorithm. The number of SNPs was reduced to those with less than 6% missing data. The core set was then compared to the entire set of Peru Uniques as well as to the group remaining after the core was removed from the entire set. Comparison was performed at summary statistics (N_e, *I*, H_o, H_e, uH_e), private alleles and Jost D_{est} as obtained in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012).

Results

Identity analysis

In the dataset of 143 trees/185 SNPs, there were nine pairs of matched samples (**Table 1**). Eight of these pairs were within the same province with the exception of INDES095 from Bagua being matched at all 185 SNPs to INDES098 from Utcubamba. A PID_{SIB} of 2.21 × 10^{-28} was obtained for the dataset of 143 trees/185 SNPs. Removal of one sample from each of the nine pairs gave a set of 134 unique samples (Peru Uniques). The Peru Uniques compared to 1,838 reference accession profiles at 175 common SNPs returned only one possible match of a putative CCN 51 to CCA015 from Condorcanqui with 171 matching loci and one mismatched locus. A PID_{SIB} of 1.862 × 10^{-30} was obtained for the dataset of 1,972 samples/175 SNPs. The average minor allele frequency over all the 562 samples is 0.261 (**Supplementary Tables 3, 4**).

Frequency analysis

The resultant frequency analysis showed that H_o was close to H_e with a very low (0.006) fixation index (**Table 2**). Using this same set of 180 SNPs, the provinces of Bagua, Condorcanqui, and Utcubamba had a low to zero fixation index (**Table 2**)

TABLE 1 Groups of matched samples in 143 cacao trees in northern Peru using 185 single nucleotide polymorphism (SNP) markers.

Group	Members (Provinces)	Number of SNPs with data	Number of matching SNPs	Number of differing SNPs	$^{a}PID_{SIB}$
1	CAP032, CAP125 (Utcubamba)	184	184	0	5.75 × 10^{-28}
2	CAP040, INDES054 (Utcubamba)	181–184	181	0	2.32 × 10^{-34}
3	CCA001, CCA021 (Condorcanqui)	183–184	183	0	7.01 × 10^{-29}
4	CCA004, CCA008 (Condorcanqui)	185	185	0	2.06 × 10^{-27}
5	CCA009, CCA010 (Condorcanqui)	185	185	0	1.83 × 10^{-34}
6	CCA016, CCA017 (Condorcanqui)	182–184	181	0	1.61 × 10^{-30}
7	INDES020, INDES023 (Bagua)	184–185	183	1	0
8	INDES021, INDES039 (Bagua)	184	184	0	5.25 × 10^{-25}
9	INDES095 (Bagua), INDES098 (Utcubamba)	185	185	0	1.51 × 10^{-31}

[a] Probability of identity that two siblings drawn at random from a population have identical genotypes (Evett and Weir, 1998; Waits et al., 2001). Estimates obtained in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012).

with slightly higher H_e than H_o in Condorcanqui but slightly higher H_o than H_e in the other two provinces. The H_{ind} in the Peru Uniques ranged from 0.056 to 0.578 with all samples being heterozygous (**Supplementary Table 5**). The lowest H_{ind} values were observed in the Utcubamba province (INDES032, H_{ind} = 0.056; INDES002, H_{ind} = 0.089). The highest H_{ind} values were observed in the Bagua (INDES070, H_{ind} = 0.578) and the Utcubamba provinces (INDES061, H_{ind} = 0.578). There was an absence of low H_{ind} (0–0.015) in Condorcanqui and Mariscal Cáceres provinces. The single sample from Jaén had a low heterozygosity (H_{ind} = 0.117; **Supplementary Table 5**).

Principal coordinate analysis

The 134 Peru Uniques were distributed across three quadrants in a linear pattern from Amelonado to Nacional but excluding close association with Criollo, Marañon and Guiana groups (**Figure 2**). There was no apparent sub-clustering of samples. The PCoA explained 23.9, 10.8, and 9.5% on the first three axes, respectively.

Phylogenetic analysis

Phylogenetic trees based on 50, 70, or 90% missing data thresholds to retain sample pairs were similar (**Supplementary Figures 1, 2**) and the tree based on 70% missing data is provided in **Figure 3**. The 134 Peru Uniques were mainly distributed between reference clusters rather than within clusters and were closest to, and arrayed along, the Nacional, Contamana, Curaray, Iquitos, Purús, and Nanay genetic groups (**Figure 3**). Two samples, CCA027 (Condorcanqui) and CAP045 (Utcubamba), were associated with the Amelonado and Criollo clusters with CAP045 being closest to the Criollo group. INDES095 from the Bagua province was an immediate sister clade to the Nanay group. Three phylogenetic clusters (Phylo A, B, C) in the Peru Uniques were found (**Figure 3**) and each cluster contained a variable number of samples from each of the three main provinces. The PhyloA cluster (represented by CAP086 from Utcubamba) contained 14 individuals (including the three samples from Mariscal Cáceres) and was positioned between the Iquitos and Purús genetic groups. The PhyloB cluster (represented by INDES064 from Utcubamba) contained 33 individuals and formed a sister clade with the Contamana/Curaray clade. The PhyloC cluster (represented by CAP107 from Utcubamba) contained 64 individuals (including the single sample from Jaén) and was positioned between the Nacional and Contamana/Curaray clades.

Group differentiation tests

The AMOVA that incorporated the 134 Peru Uniques as a unit group partitioned 54.5% within genetic clusters and 45.5% among genetic clusters (**Supplementary Table 6**). The genetic differentiation (F_{ST} = 0.455) was significant (P = 0.001) (**Supplementary Table 6**). The F_{ST} among pairwise groups (**Supplementary Table 7**) indicated that the set of Peru Uniques was closest to the group of Amelonado/Nacional hybrids (0.126) and then to the Nacional cluster (0.155). The Jost D_{est} measure of group differentiation among the reference groups were all significant (P = 0.001, 0.002) with maximal values of 0.626 (Amelonado vs. Criollo) and minimal values of 0.089 (Amelonado vs. Amelonado/Criollo) and 0.098 (Amelonado/Nacional vs. Amelonado/Criollo) (**Figure 4** and **Supplementary Tables 8, 9**). Private alleles were only present from two SNPs (TcSNP0097, TcSNP1158) in the Contamana group.

Jost D_{est} measure was obtained from a dataset of 136 SNPs for which missing data was less than 6% in any of the grouped samples. The D_{est} values were significant (P = 0.001) for all pairwise comparisons involving the three Peruvian provinces and any of the 12 reference groups with maximal D_{est} (0.347) recorded for Criollo vs. Condorcanqui (**Supplementary Table 8**). The reference group Iquitos was closest to Condorcanqui (D_{est} = 0.098) whereas Nacional was closest to Bagua (0.075) and Utcubamba (0.066). The three provinces were also close to the set of Amelonado/Nacional hybrids (D_{est} = 0.070 – 0.076). Among the three provinces, significant and low D_{est} values were obtained for Bagua vs. Condorcanqui (D_{est} = 0.016; P = 0.006) and Condorcanqui vs.

TABLE 2 Descriptive genetic statistics for set of unique cacao and its core collection in north Peru with 180 SNPs.

	N_e	I	H_o	H_e	uH_e	FIS	F
Peru Uniques (Entire, E) (n = 134)	1.563 ± 0.022	0.507 ± 0.012	0.340 ± 0.012	0.336 ± 0.010	0.337 ± 0.010	0.006 ± 0.014	−0.012
Bagua(n = 16)	1.579 ± 0.023	0.509 ± 0.013	0.389 ± 0.015	0.339 ± 0.011	0.350 ± 0.011	−0.126 ± 0.018	−0.147
Condorcanqui (n = 24)	1.581 ± 0.023	0.509 ± 0.013	0.319 ± 0.012	0.340 ± 0.011	0.347 ± 0.011	0.065 ± 0.017	0.061
Utcubamba(n = 91)	1.526 ± 0.022	0.486 ± 0.013	0.336 ± 0.013	0.320 ± 0.010	0.321 ± 0.010	−0.022 ± 0.016	−0.052
Core(C; n = 53)	1.594 ± 0.021	0.526 ± 0.011	0.334 ± 0.011	0.351 ± 0.009	0.354 ± 0.009	0.056 ± 0.015	0.049
Entire-Core(E-C; n = 81)	1.540 ± 0.022	0.489 ± 0.013	0.344 ± 0.013	0.323 ± 0.011	0.325 ± 0.011	−0.039 ± 0.015	−0.064

Values are mean ± standard error. N_e = number of effective alleles per locus; I = Shannon's Information Index; H_o = observed heterozygosity; H_e = expected heterozygosity; uH_e = unbiased expected heterozygosity; FIS = Average fixation index of loci; F = fixation index for population. Estimates obtained in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012).

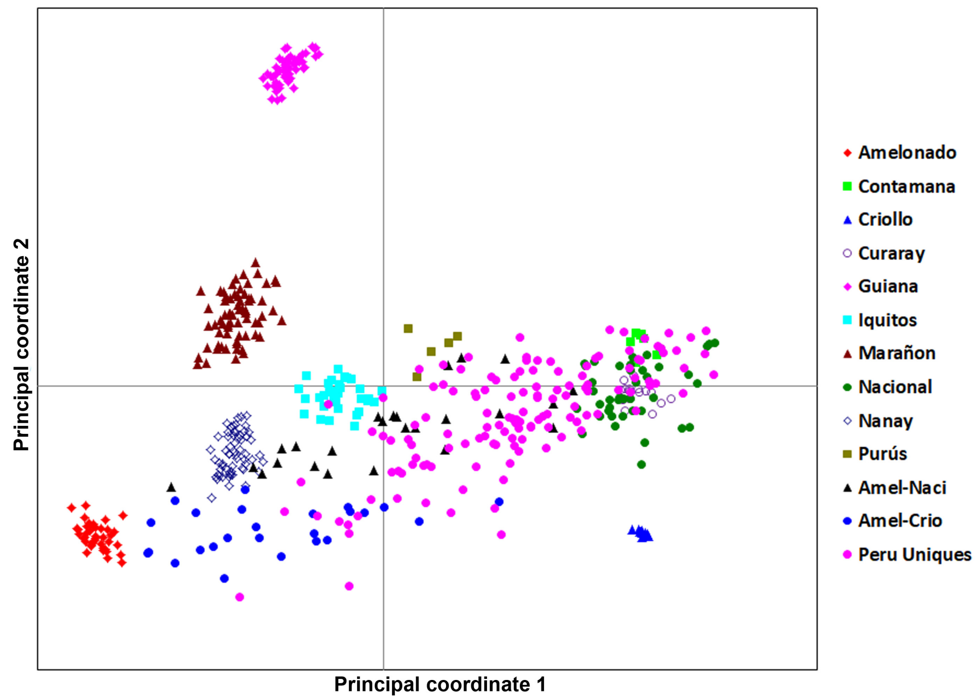


FIGURE 2

A principal coordinate analysis 2D-scatter plot of 134 Peru Uniques and 390 reference accessions using 170 SNP genetic data. The first and second axes explained 23.86 and 10.80% of the variation, respectively.

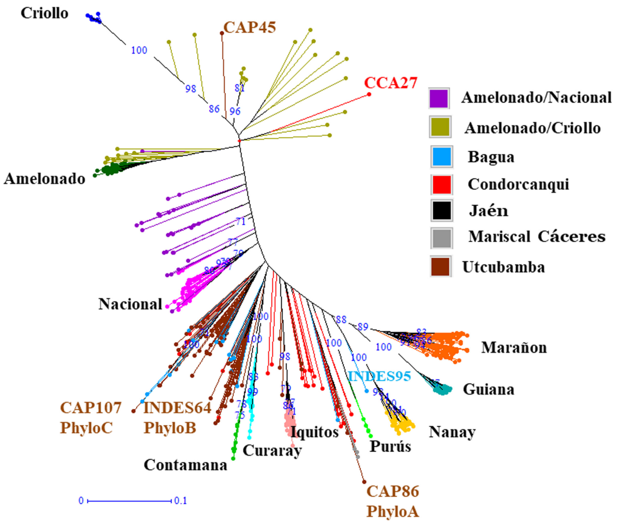


FIGURE 3

Phylogram (based on 70% missing data) of unique cacao samples collected from Northern Peru (134 samples) and 390 reference accessions using 170 single nucleotide polymorphisms. Three phylogenetic clusters (PhyloA-C) from the samples from north Peru are indicated together with a representative sample. All three representative samples are from Utcubamba and each cluster also contain samples from Bagua and Condorcanqui. Three other samples from north Peru are indicated – CAP45 (Utcubamba), CCA27 (Condorcanqui) and INDES95 (Bagua). The three samples from Mariscal Cáceres are in PhyloA and the one sample from Jaén is in PhyloC. ≥ 70% bootstrap values are displayed.

Utcubamba (D_{est} = 0.016; P = 0.006) but Bagua vs. Utcubamba was low and non-significant (D_{est} = 0.001; P = 0.285). Phylogenetic clusters of the 10 population groups of Motamayor et al. (2008), two reference mixed groups (Amelonado/Nacional and Amelonado/Criollo) and the three clades (Phylo A, B, C) in the Peruvian dataset were all significantly different (P = 0.001) from each other (**Supplementary Table 9**). The PhyloB and PhyloC clades were closest to each other (D_{est} = 0.037) while the PhyloA and PhyloC clades were furthest from other Peruvian samples (D_{est} = 0.144). The three clades in the Peruvian dataset were all close to the Amelonado/Nacional group (D_{est} = 0.094–0.097). The PhyloA clade was closest to the Iquitos (D_{est} = 0.082) and the Amelonado/Criollo (D_{est} = 0.094) reference groups. The PhyloB clade was closest to the Amelonado/Nacional (D_{est} = 0.094) reference group. The PhyloC clade was closest to the Nacional (D_{est} = 0.063) and the Amelonado/Nacional (D_{est} = 0.095) reference groups.

Ancestry analysis

At K = 10, the reference populations were all resolved into the expected groupings of Motamayor et al. (2008) and the Amelonado/Nacional and Amelonado/Criollo hybrids presented the ancestry profile to match their expected founder populations (**Figure 5A**). However, the optimal K value by Evanno's method was for 11 populations. At K = 11, the 10 reference populations were resolved, but one population in each run was split into two distinct groups. The population that was split was inconsistent and occurred in the Amelonado (once), Contamana (twice), Criollo (thrice), Curaray (twice), Guiana (once), and Marañon (once) of the 10 iterations. An example is presented in **Figure 5B**. Nonetheless, the Peru Uniques at both K = 10 and K = 11 in the initial analysis exhibited a strong and frequent Nacional background (**Figures 5A,B**) which was supported by the more stringent analysis at K = 10 (**Supplementary Tables 10, 11**).

The 134 Peru Uniques were all admixed with contributions from 2 to 7 genetic groups and with a mixture of three groups being the most frequent (**Figure 5C**). Apart from Nacional, at least 50% ancestry was present from Amelonado (CCA27, Condorcanqui), Contamana (CAP107, Utcubamba; INDES70, Bagua), Criollo (CAP45; Utcubamba) and Iquitos (CAP37, INDES18, INDES62: Utcubamba; CCA16, Condorcanqui). The combination of only Amelonado with Criollo ancestry was only found in CAP45 and CCA27. The single sample collected from Jaén had 85% Nacional, 8% Curaray, and 5% Iquitos ancestral background. The three samples from Mariscal Cáceres had a common background of Amelonado, Criollo, and Iquitos. However, apart from the actual contributions being different, two of these (INDES 101, 106) were higher in Iquitos (41%) ancestry whereas the other (INDES112) had Nanay as the major component (37%). A set of 49 samples combined both Criollo and Nacional ancestry (each at a minimum of 10%) with the majority (36) coming from the Utcubamba province and the remainder from the Bagua (8) and Condorcanqui (3) provinces. Sixteen samples lacked Nacional ancestry and contained instead an Amelonado/Criollo background with other groups except for INDES95 from Bagua which was mixed with Nanay (44%), Iquitos (36%), and Curaray (13%). The proportion of cacao trees with at least 25% Nacional ancestry was highest in Utcubamba (91.1%; 82 of 90), then Bagua (75%; 12 of 16) with the lowest occurrence in Condorcanqui (58.3%, 14 of 24). However, only the comparison of Condorcanqui to Utcubamba was significantly different (P = 0.0003; **Supplementary Table 12**).

Core collection identification

A set of 53 samples were identified as a core collection from the 134 Peru Uniques and 182 SNP loci (**Supplementary Table 13**). Statistical measures in the core were higher than that of the entire set and the entire set without the core except for H_o which showed the reverse trend (**Table 2**). However, private alleles were lacking and Jost D_{est} was non-significant being estimated as 0 (P = 0.993) and 0.001 (P = 0.177) for the aforementioned two comparisons.

Discussion

Examining the genetic diversity and ancestry of cacao from its centre of diversity is essential to better understand its population structure and the judicious conservation and cultivation of native varieties. In this study 143 cacao trees from north-western Peru (the majority from the Amazonas region) were SNP genotyped *via* 185 informative SNPs to examine their genetic diversity and ancestry. Overall, the findings indicated that the samples had moderate gene diversity (H_e = 0.336) and shared ancestry with the Nacional, Amelonado, Iquitos and Criollo groups. The 143 samples had few matching duplicates (nine groups with two members each) and these were usually within a province rather than across provinces. This internal duplicate matching was lower than that recorded for cacao collected in Belize (Motilal et al., 2010) and for farm selections in Dominica (Gopaulchan et al., 2019), Dominican Republic (Boza et al., 2013), Hawaii (Nagai et al., 2009), Nicaragua (Trognitz et al., 2011), the Huallaga and Ucayali valleys in Peru (Zhang et al., 2006a), Puerto Rico (Cosme et al., 2016) but higher than that reported in one farm in Jamaica (Lindo et al., 2018), Vietnam (Everaert et al., 2017) or for the ICS and TRD accessions in Trinidad (Johnson et al., 2009). Furthermore, an absence of duplicates was reported for 164 trees in Bolivia (Zhang et al., 2012), 93 trees in Tumaco, Colombia (Yacenia Morillo et al., 2014), for 53 trees in Sulawesi, Indonesia (Dinarti et al., 2015), for 220 trees in the Juanjui province of the Huallaga valley, Peru (Zhang et al., 2011), and for 109 trees in Uganda (Gopaulchan et al., 2019). A set of 134 Peru Uniques was obtained after removal of duplicate samples and only one sample (CCA015; Condorcanqui) matched to an external reference variety with a very low PID_{SIB} (1.862×10^{-30}) in the identity analysis dataset. The internal and external match

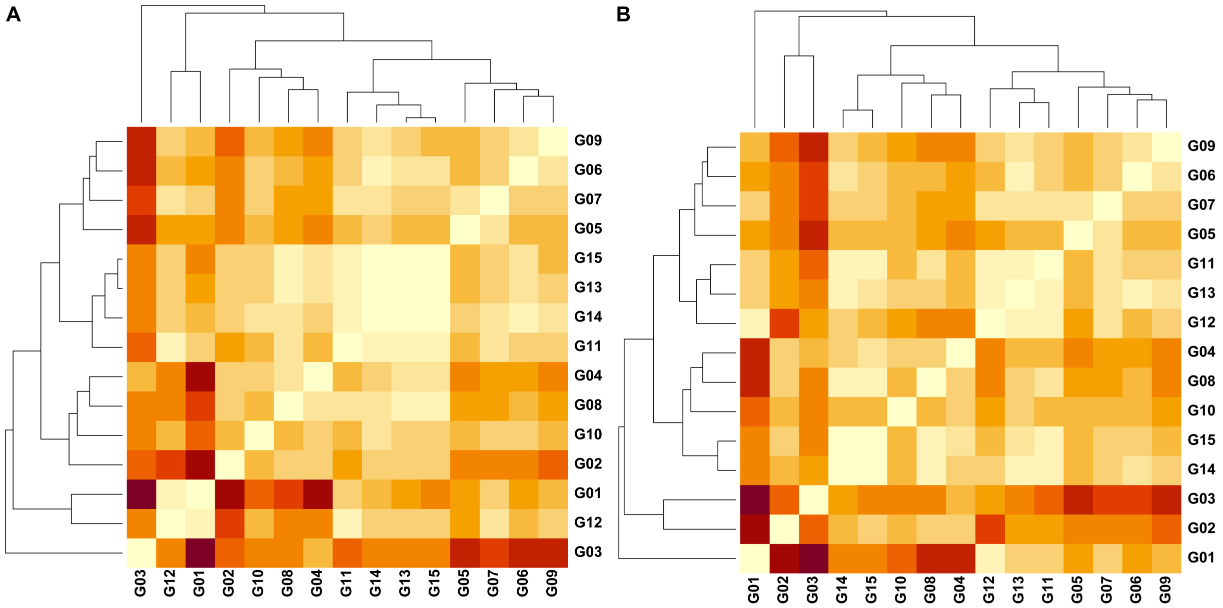


FIGURE 4

Heatmaps representing pairwise Jost differentiation indices (D_{est}) (light = 0.00/dark = 0.626) among 15 groups of cacao samples based on the genetic variation of 136 SNPs examining **(A)** three regions and **(B)** three phylogenetic clusters in north Peru. Groups – G1 (Amelonado; *n* = 40), G2 (Contamana; *n* = 8), G3 (Criollo; *n* = 12), G4 (Curaray; *n* = 17), G5 (Guiana; *n* = 56), G6 (Iquitos; *n* = 32), G7 (Marañon; *n* = 70), G8 (Nacional; *n* = 40), G9 (Nanay; *n* = 60), G10 (Purús; *n* = 5), G11 (Amelonado/Nacional; *n* = 25), G12 (Amelonado/Criollo; *n* = 25), $G13_A$ (Bagua; *n* = 16), $G14_A$ (Condorcanqui; *n* = 24), $G15_A$ (Utcubamba; *n* = 91), $G13_B$ (PhyloA; *n* = 14), $G14_B$ (PhyloB; *n* = 33), $G15_B$ (PhyloC; *n* = 64). D_{est} values obtained in GenAlEx v6.502 (Peakall and Smouse, 2006, 2012).

analyses indicated that the cacao samples collected in north Peru were generally distinct and unique. This is promising for maintaining relic diversity, identifying genotypes best suited to local conditions and maintaining the distinctiveness of the Peruvian fine aroma cacao industry. The few duplicate trees may represent very closely related varieties that are unable to be resolved with the SNP panel in this study. If not, then these duplicated samples may represent clonal propagated material that was disseminated in earlier years. The presence of a sample similar to a putative CCN 51 supports the latter view and represents a cautionary note for north Peru. However, the ancestry profile of the sample in north Peru (CCA015) was different from that reported in Boza et al. (2014) suggesting that the CCN 51 may have been a mislabelled reference accession. CCA015, while probably not CCN 51, represents an example of a sample without Nacional but with Criollo ancestry which will contribute to the fine aroma designation.

The low or zero fixation indices are indicative of the absence of inbreeding. This supports the use of crosses between trees sampled from the Peruvian Amazon region for genetic improvement. A moderate level of gene diversity was observed (H_e = 0.32–0.34) for the Peru Uniques as well as in the Bagua, Condorcanqui, and Utcubamba provinces. This was similar to on-farm cacao in Dominica (H_e = 0.320; Gopaulchan et al., 2020), in Honduras and Nicaragua (H_e = 0.367; Lukman et al., 2014), and Uganda (H_e = 0.332; Gopaulchan et al., 2019) but higher than in Colombia (H_e = 0.28; Yacenia Morillo et al., 2014), Ghana (H_e = 0.245; Padi et al., 2015), and Chuncho cacao from the La Convención province in south Peru (H_e = 0.230; Céspedes-Del Pozo et al., 2018). The H_e of cacao in north Peru was lower than that reported in Bolivia (H_e = 0.56; Zhang et al., 2012), Cameroon (H_e = 0.50; Efombagn et al., 2008), of the Juanjui province of San Martin in north Peru (H_e = 0.741; Zhang et al., 2011) and that of Ecuador (H_e = 0.496; Loor Solorzano et al., 2009). The moderate H_e observed in this study is probably reflective of the lack of imported varieties to give rise to differential hybrid material. The higher H_e reported above may also have been due in part to the use of microsatellites in those studies.

Estimates of H_{ind} revealed that the majority of the collected samples were heterozygous with few highly homozygous trees. In contrast, Lerceteau et al. (1997) reported a high level of homozygous trees in two old plantations (80–100 years) in Ecuador. This indicated a low incidence of inbreed individuals in the current study and the presence of good cross-compatibilities among a greater number of founder individuals in the current study. Highly heterozygous samples should be assessed for vigour and productivity. The eleven samples with low heterozygosity should be assessed for self-compatibility toward obtaining pure lines for breeding purposes. Differential phenotypes selected from these two

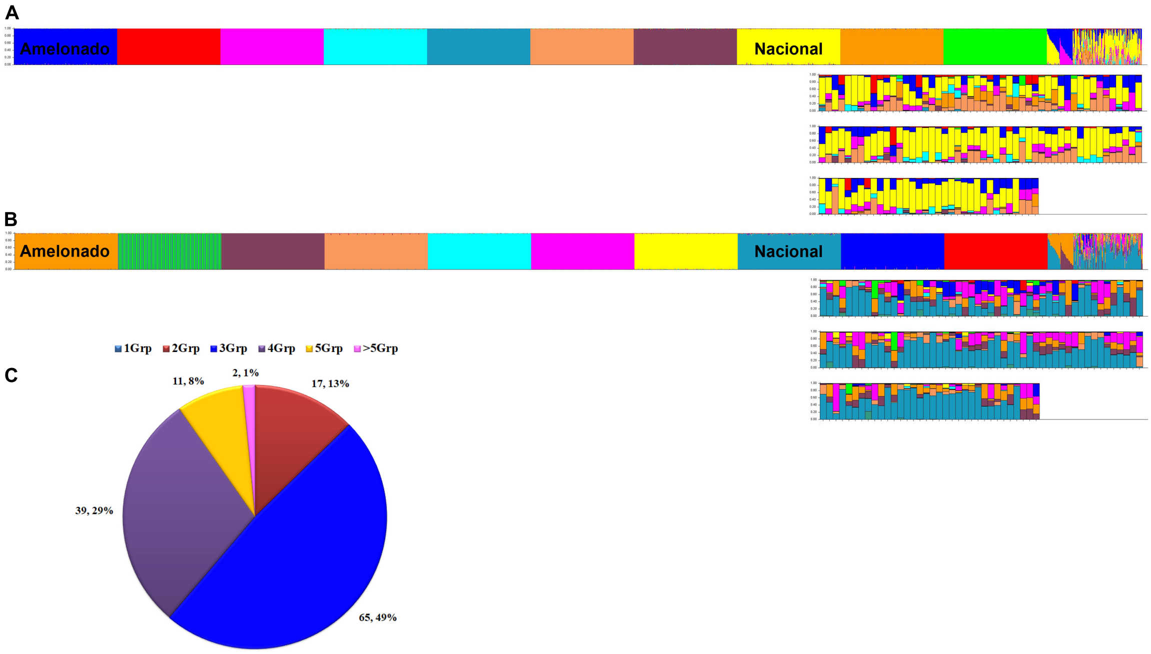


FIGURE 5
Ancestry of 134 unique cacao samples (Peru Uniques) collected from northern Peru. Ancestry at $K = 10$ **(A)** and $K = 11$ **(B)** using 154 SNPs obtained from STRUCTURE (Pritchard et al., 2000) output using model based on 100,000 burnins, 200,000 Markov Chain Monte Carlo (MCMC) simulations, admixture ancestry model and independent allele frequencies. Samples are arranged as 10 reference populations (Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay, Purús), Amelonado/Nacional, Amelonado/Criollo and Peru Uniques. Distribution of admixture classes **(C)** in Peru Uniques from 1 to more than five genetic groups (Grp) obtained from STRUCTURE (Pritchard et al., 2000) output using model based on 170 SNPs, 300,000 burnins, 600,000 MCMC simulations, admixture ancestry model and independent allele frequencies.

groups may be useful to find QTL (quantitative trait locus) for tree breeding purposes. Although the samples had mainly heterozygous individuals, the Shannon Index of diversity was similar to that reported in Dominica (Gopaulchan et al., 2020) and Uganda (Gopaulchan et al., 2019) but lower than in Honduras and Nicaragua (Ji et al., 2013), and in Indonesia (Lukman et al., 2014). The samples from north Peru were therefore lower in diversity and probably reflects the lower occurrence of introduced germplasm from other countries.

The PCoA revealed an underlying pattern of mixed types between Amelonado and Nacional groups that was supported by the phylogenetic, group differentiation and ancestry analyses. Three distinct phylogenetic clusters were present in the collected germplasm and supported by D_{est} statistics. AMOVA, F_{ST}, and D_{est} analyses supported the distinction of the Peru Uniques from the reference groups and the provinces of Bagua, Condorcanqui and Utcubamba from the reference groups. However, the provinces of Bagua and Utcubamba were similar to each other. This differed from Oliva-Cruz (2020) who found that these two provinces differed in ecotype composition. Yet, the proportion of Nacional trees from the current study was similar between these two provinces. These results and the identity analyses suggest that the sampled germplasm in north Peru contained unique multilocus profiles with possible inter-provincial differentiation and greater similarity between the Bagua and Utcubamba provinces. The province of Condorcanqui is recommended for further collection to verify its difference. Likewise, the province of Bagua was represented by 16 samples and increasing the sample size would allow for better resolution of inter-provincial differentiation. Bulking of cacao samples for fermentation or marketing purposes could be undertaken for the provinces of Bagua and Utcubamba. A similar recommendation for bulking across regions was obtained for Dominica (Gopaulchan et al., 2020). Further refinement could be achieved by propagating and maintaining the three phylogenetic clusters as distinct units provided that their sensory profiles are different. Additional collection and SNP genotyping to ascertain the frequency and distribution of these three clades in north Peru should be undertaken.

The clade PhyloC with 64 members was a good candidate for a new genetic group based on the phylogram (**Figure 3**). Initial population modelling in STRUCTURE (Pritchard et al., 2000) and assessed with the method of Evanno et al. (2005) fitted 11 groups. However, this was at the expense of splitting an accepted genetic cluster into two distinct groups instead of identifying PhyloC as a separate group. Furthermore, the two mixed reference groups (Amelonado/Nacional and Amelonado/Criollo) were also differentiated from D_{est} estimates from all other groups indicative that samples just need to be in groups rather than true populations to have differing estimators of genetic differentiation. None of the three test clades had any private alleles that could have supported the presence of a different genetic cluster. The results were therefore interpreted as PhyloA, PhyloB, and PhyloC being better fitted as clades of germplasm with hybrid ancestry from the genetic grouping of Motamayor et al. (2008). Hence, the collected samples from north Peru were mainly unique admixed cacao trees but did not comprise a novel genetic cluster and did not contain a subset that could be a novel group.

New populations in cacao were reported for Bolivia (Zhang et al., 2012), Colombia (Osorio-Guarín et al., 2017), and Peru (Céspedes-Del Pozo et al., 2018). However, these reports of new populations may be tentative due to limitations in each study. The Beni population in Bolivia was shown to be distinct from the Ucayali population from F_{ST} values (Zhang et al., 2012). However, an ancestry plot as well as a phylogenetic tree were not provided and the possibility of a sister clade to the Ucayali population cannot be ruled out. The Ucayali population contains members of the SCA accessions (Zhang et al., 2011) which belong to the Contamana cluster (Motamayor et al., 2008). Hence, the Beni population could be like clade PhyloB which was significantly different by D_{est} statistic from Contamana but was not a unique group. Furthermore, the PCoA study of Zhang et al. (2012) was limited by the few representative members of the accepted 10 genetic clusters and was probably lacking Purús members which may have resulted in an artificial separation of the Beni germplasm from the reference accessions.

Osorio-Guarín et al. (2017) indicated that two new cacao groups in Colombia were present based primarily on their ancestry result. However, examination of their graph revealed that the Iquitos and Nanay genetic clusters were not resolved from each other and the Curaray population was composed of three groups including Contamana and Amelonado. This suggests that the new groups were at the expense of established populations and further modelling is required to firmly establish whether these are new genetic clusters, subgroups of existing populations or sister clades of related germplasm. Céspedes-Del Pozo et al. (2018) reported that the Chuncho cacao from the La Convención province in Cusco, Peru was a distinct genetic cluster even from Contamana. These authors found very close genetic distances (0.06–0.07) of Chuncho to the Beni, Madre de Dios and Ucayali groups similar to the close D_{est} values for the Peru Uniques and the clades PhyloA, Phylob and PhyloC to the Nacional and Iquitos genetic groups in the current study. Furthermore, the PCoA plot of Céspedes-Del Pozo et al. (2018) apparently did not employ members from six known genetic groups including Nacional and Curaray thereby compromising the suggested distinct clustering of Chuncho cacao. In addition, some members of the Ucayali/Urubamba which are likely members of the Contamana cluster were dispersed among the Chuncho samples. Examination of the ancestry graphs of Céspedes-Del Pozo et al. (2018) indicated that two genetic groups, likely Iquitos and Nanay, were unresolved as in Osorio-Guarín et al. (2017). The allocation of Chuncho to a new group may therefore need further validation.

The fit to the 10 genetic groups of Motamayor et al. (2008) agreed with the close D_{est} values to Iquitos, Nacional, Amelonado/Nacional and Amelonado/Criollo groups. As a unit group, the Peru Uniques was closest to the set of Amelonado/Nacional mixed references with the Bagua, Condorcanqui and Utcubamba provinces being closest to Nacional and Amelonado/Nacional groups. A similar result was returned for test clades PhyloB and PhyloC whereas PhyloA was closest to the Iquitos group. Actual ancestry estimates supported the predominance of Nacional ancestry variably mixed mainly with Amelonado and Iquitos with additional contributions from Contamana, Criollo and Nanay groups. Only two Amelonado/Criollo admixed samples were found in north Peru, moderate Criollo ($\geq$30%) ancestry was found in only six samples and only 16 samples lacked Nacional ancestry. This suggests that the fine aroma of cacao in north Peru is likely due to the Nacional background. However, the Condorcanqui province had a lower occurrence of Nacional members and it may be worthwhile to rejuvenate or infill farms with accessions having high Nacional ancestry already present in this region to ensure that the fine aroma status is maintained. The 16 samples lacking Nacional ancestry should be revisited and assessed for disease, productivity and flavour traits. If they prove to have superior traits including valuable or marketable flavour attributes, these samples can be cloned and maintained for breeding purposes. However, if acceptable trait combinations are lacking, these trees should not be clonally propagated or used to obtain open-pollinated seeds for distribution to farmers. This will help to maintain the fine aroma designation. Similarly, the set of 49 samples that contained both Criollo and Nacional ancestry should be examined for their flavour profile. If a distinctive flavour profile is found, this group can be clonally propagated and distributed to farmers. Self- and cross-compatibilities should be ascertained prior to distribution to identify the best possible mix to achieve fruit set on farms.

The genetic diversity of the fine aroma cacao in north Peru could be adequately represented by a set of 53 samples. The 143 sampled trees of this study have been clonally propagated as rooted cuttings and maintained as three different germplasm collections in an altitudinal gradient in the Utcubamba province (420 masl, 779620.6S, 9363856.4W; 480 masl, 792305.8S, 9364081.9W; 950 masl, 801491.0S, 9364914.0W). This collection will be expanded as additional genotyping is obtained on germplasm from future field collections. Furthermore, phenotyping of the three germplasm collections would provide information on phenotypic diversity that can be used to complement the genetic diversity of the set of 53 accessions and hence obtain a best core collection and a working collection. The core collection should be safeguarded by having an internal safety duplication where each accession is represented by at least five clonal copies and by having replicates of the core collection at different sites within the Peruvian Amazonas region. This would facilitate access to budwood for propagation to resupply farms with best local material to maintain the fine aroma status of cacao in north Peru.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

Author contributions

DB, MC, and MO conceived the idea and acquired the funding for the research and collecting expedition. LM and AM curated the data and selected the reference accessions. LM conducted the data analysis. DB, MC, AM, and LM contributed to the first draft of the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

Funding

This study was supported by the Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (FONDECYT) funded by the Project through the Contract N 026-2016 "Círculo de Investigación para la Innovación y el fortalecimiento de la cadena de valor del cacao nativo fino de aroma en la zona nor oriental del Perú (CINCACAO)". This study was also partially funded by the Project through the Contract 030-2018-FONDECYT-BM-IADT-MU and 142-2018-FONDECYT-BM-IADT-MU.

Acknowledgments

We are most grateful to Jani Mendoza, Rosmery Robles, Jhordy Perez, and Daniel Tineo for their technical and logistical assistance. We also thank to Clelia Jima Chamiquit and Stefhany Valdeiglesias Ichillumpa for their translations to Awajun and Quechua languages, respectively.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.895056/full#supplementary-material

References

Alverson, W. S., Whitlock, W. A., Nyffeler, R., Bayer, C., and Baum, D. (1999). Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *Am. J. Bot.* 86, 1474–1486. doi: 10.2307/2656928

Arevalo-Gardini, E., Meinhardt, L. W., Zuñiga, L. C., Arévalo-Gardni, J., Motilal, L., and Zhang, D. (2019). Genetic identity and origin of "Piura Porcelana"—a fine-flavored traditional variety of cacao (*Theobroma cacao*) from the Peruvian Amazon. *Tree Genet. Genomes* 15:11. doi: 10.1007/s11295-019-1316-y

Bartley, B. G. D. (2001). Refractario—an explanation of the meaning of the term and its relationship to the introductions from Ecuador in 1937. *Ingenic Newslett.* 6, 10–15.

Bartley, B. G. D. (2005). *The Genetic Diversity of Cacao and Its Utilization*. Cambridge MA: CABI Publishing.doi: 10.1079/9780851996196.0000

Bayer, C., Fay, M. F., de Bruijn, A. Y., Savolainen, V., Morton, C. M., Kubitzki, K., et al. (1999). Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: a combined analysis of plastid *atp*B and *rbc*L DNA sequences. *Bot. J. Linn.* 129, 267–303. doi: 10.1111/j.1095-8339.1999.tb00505.x

Beg, M. S., Ahmad, S., Jan, K., and Bashir, K. (2017). Status, supply chain and processing of cocoa–A review. *Trends Food Sci. Technol.* 66, 108–116. doi: 10.1016/j.tifs.2017.06.007

Boza, E. J., Irish, B. M., Meerow, A. W., Tondo, C. L., Rodríguez, O. A., Ventura-López, M., et al. (2013). Genetic diversity, conservation, and utilization of *Theobroma cacao* L.: genetic resources in the Dominican Republic. *Genet. Resour. Crop Evol.* 60, 605–619. doi: 10.1007/s10722-012-9860-4

Boza, E. J., Motamayor, J. C., Amores, F. M., Cedeño-Amador, S., Tondo, C. L., Livingstone, D. S., et al. (2014). Genetic characterization of the cacao cultivar

CCN 51: its impact and significance on global cacao improvement and production. *J. Amer. Soc. Hort. Sci.* 139, 219–229. doi: 10.21273/JASHS.139.2.219

Céspedes-Del Pozo, W. H., Blas-Sevillano, R., Zhang, D., and University students (2018). "Assessing genetic diversity of cacao (*Theobroma cacao* L.) nativo Chuncho in La Convención, Cusco-Perú," in *Proceedings of the International Symposium on Cocoa Research (ISCR)*, Lima.

Chia-Wong, J. A., Márquez-Dávila, K. J., Cárdenas-Salazar, H., Hurtado-Gonzales, O. P., Huaman-Camacho, T., Céspedes-Del-Poso, W., et al. (2018). "Avances en el estudio de las bases genéticas y organolépticas del cacao fino o de aroma en el Perú," in *Proceedings of the International Symposium on Cocoa Research (ISCR)*, Lima.

Cosme, S., Cuevas, H. E., Zhang, D., Oleksyk, T. K., and Irish, B. M. (2016). Genetic diversity of naturalized cacao (*Theobroma cacao* L.) in Puerto Rico. *Tree Genet. Genomes* 12:88. doi: 10.1007/s11295-016-1045-4

De la Fuente, L. (2018). *El cacao, tesoro de la Amazonía*. Lima: Fondo Editorial USIL.

Dinarti, D., Susilo, A. W., Meinhardt, L. W., Ji, K., Motilal, L. A., Mischke, S., et al. (2015). Genetic diversity and parentage in farmer selections of cacao from Southern Sulawesi, Indonesia revealed by microsatellite markers. *Breed. Sci.* 65, 438–446. doi: 10.1270/jsbbs.65.438

Efombagn, I. B. M., Motamayor, J. C., Sounigo, O., Eskes, A. B., Nyassé, S., Cilas, C., et al. (2008). Genetic diversity and structure of farm and GenBank accessions of cacao (*Theobroma cacao* L.) in Cameroon revealed by microsatellite markers. *Tree Genet. Genomes* 4, 821–831. doi: 10.1007/s11295-008-0155-z

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

Everaert, H., De Wever, J., Tang, T. K. H., Vu, T. L. A., Maebe, K., Rottiers, H., et al. (2020). Genetic classification of Vietnamese cacao cultivars assessed by SNP and SSR markers. *Tree Genet. Genomes* 16:43. doi: 10.1007/s11295-020-01439-x

Everaert, H., Rottiers, H., Pham, P. H. D., Ha, L. T. V., Nguyen, T. P. D., Tran, P. D., et al. (2017). Molecular characterization of Vietnamese cocoa genotypes (*Theobroma cacao* L.) using microsatellite markers. *Tree Genet. Genomes* 13:99. doi: 10.1007/s11295-017-1180-6

Evett, I. W., and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates.

Fang, W., Meinhardt, L. W., Mischke, S., Bellato, C. M., Motilal, L., and Zhang, D. (2014). Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. *J. Agric. Food Chem.* 62, 481–487. doi: 10.1021/jf404402v

FAO (2022). *FAOSTAT Crops and Livestock Products. License: CC BY-NC-SA 3.0 IGO*. Available online at: https://www.fao.org/faostat/en/#data/QCL (accessed May 28, 2022).

Goenaga, R., Irizarry, H., and Irish, B. (2009). TARS Series of Cacao Germplasm Selections. *HortScience* 44, 826–827. doi: 10.21273/HORTSCI.44.3.826

Gopaulchan, D., Motilal, L. A., Bekele, F. L., Clause, S., Ariko, J. O., Ejang, H. P., et al. (2019). Morphological and genetic diversity of cacao (*Theobroma cacao* L.) in Uganda. *Physiol. Mol. Biol. Plants.* 25, 361–375. doi: 10.1007/s12298-018-0632-2

Gopaulchan, D., Motilal, L. A., Kalloo, R. K., Mahabir, A., Moses, M., Joseph, F., et al. (2020). Genetic diversity and ancestry of cacao (*Theobroma cacao* L.) in Dominica revealed by single nucleotide polymorphism markers. *Genome* 63, 583–595. doi: 10.1139/gen-2019-0214

Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual [INDECOPI] (2016). *Denominación de Origen Cacao Amazonas Perú*. Lima: INDECOPI.

Instituto Nacional de Estadística e Informática [INEI] (2012). *Características de la Unidad Agropecuaria in IV Censos Nacional Agropecuario*. Lima: INEI.

International Cocoa Organization [ICCO] (2021). *Fine or Flavour Cocoa*. New Delhi: ICCO.

Ji, K., Zhang, D., Motilal, L. A., Boccara, M., Lachenaud, P., and Meinhardt, L. W. (2013). Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) marker. *Genet. Resour. Crop. Evol.* 60, 441–453. doi: 10.1007/s10722-012-9847-1

Johnson, E. S., Bekele, F., Brown, S., Song, Q., Zhang, D., Meinhardt, L. W., et al. (2009). Population structure and genetic diversity of the Trinitario cacao (*Theobroma cacao* L.) from Trinidad and Tobago. *Crop Sci.* 49, 564–572. doi: 10.2135/cropsci2008.03.0128

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. (2009). D vs, GST: response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Mol. Ecol.* 18, 2088–2091. doi: 10.1111/j.1365-294X.2009.04186.x

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* 16, 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Laliberté, B., End, M., Cryer, N., Daymond, A., Engels, J., Eskes, A. B., et al. (2018). "Conserving and exploiting cocoa genetic resources: The key challenges," in *Achieving Sustainable Cultivation of Oil Palm*, ed. D. Burleigh (Cambridge: Science Publishing), 19–46. doi:

Lerceteau, E., Quiroz, J., Soria, J., Flipo, S., Pe'tiard, V., and Crouzilat, D. (1997). Genetic differentiation among Ecuadorian *Theobroma cacao* L. accessions using DNA and morphological analyses. *Euphytica* 95, 77–87. doi: 10.1023/A:1002993415875

Lindo, A. A., Robinson, D. E., Tennant, P. F., Meinhardt, L. W., and Zhang, D. (2018). Molecular characterization of cacao (*Theobroma cacao*) germplasm from Jamaica using single nucleotide polymorphism (SNP) markers. *Trop. Plant Biol.* 11, 93–106. doi: 10.1007/s12042-018-9203-5

Livingstone, D. S., Freeman, B., Motamayor, J. C., Schnell, R. J., Royaert, S., Takrama, J., et al. (2012). Optimization of a SNP assay for genotyping *Theobroma cacao* under field conditions. *Mol. Breeding* 30, 33–52. doi: 10.1007/s11032-11011-19596-11034

Loor Solorzano, R. G., Risterucci, A. M., Courtois, B., Fouet, O., Jeanneau, M., Rosenquist, E., et al. (2009). Tracing the native ancestors of modern *Theobroma cacao* L. population in Ecuador. *Tree Genet. Genomes* 5, 421–433. doi: 10.1007/s11295-008-0196-3

Lukman., Zhang, D., Susilo, A. W., Dinarti, D., Bailey, B., Mischke, S., et al. (2014). Genetic identity, ancestry and parentage in farmer selections of cacao from Aceh, Indonesia revealed by single nucleotide polymorphism (SNP) markers. *Trop. Plant Biol.* 7, 133–144. doi: 10.1007/s12042-014-9144-6

Mahabir, A., Motilal, L. A., Gopaulchan, D., Ramkissoon, S., Sankar, A., and Umaharan, P. (2020). Development of a core SNP panel for cacao (*Theobroma cacao* L.) identity analysis. *Genome* 63, 103–114. doi: 10.1139/gen-2019-0071

MedCalc Software bvba (2013). *MedCalc Statistical Software version 12.7.7*. Ostend: MedCalc Software bvba.

Mejía, A., Meza, G., Espichan, F., Mogrovejo, J., and Rojas, R. (2021). Chemical and sensory profiles of Peruvian native cocoas and chocolates from the Bagua and Quillabamba regions. *Food Sci. Technol.* 41, 576–582. doi: 10.1590/fst.08020

Motamayor, J. C., Lachenaud, P., da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., et al. (2008). Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS One* 3:e3311. doi: 10.1371/journal.pone.0003311

Motamayor, J. C., and Lanaud, C. (2002). "Molecular analysis of the origin and domestication of *Theobroma cacao* L," in *Managing Plant Genetic Diversity*, eds J. M. M. Engels, V. Ramanatha Rao, A. H. D. Brown, and M. T. Jackson (Oxon: CABI Publishing). doi: 10.1079/9780851995229.0077

Motilal, L. A., Sankar, A., Gopaulchan, D., and Umaharan, P. (2017). "Cocoa," in *Biotechnology of plantation crops*, eds P. Chowdappa, A. Karun, M. K. Rajesh, and S. V. Ramesh (New Delhi: Daya Publishing House), 313–354.

Motilal, L. A., Zhang, D., Umaharan, P., Mischke, S., Boccara, M., and Pinney, S. (2009). Increasing accuracy and throughput in large-scale microsatellite fingerprinting of cacao field germplasm collections. *Trop. Plant Biol.* 2, 23–37. doi: 10.1007/s12042-008-9016-z

Motilal, L. A., Zhang, D., Umaharan, P., Mischke, S., Mooleedhar, V., and Meinhardt, L. W. (2010). The relic Criollo cacao in Belize - Genetic diversity and relationship with Trinitario and other cacao clones held in the International Cocoa Genebank. *Trinidad. Plant Genet. Resour.* 8, 106–115. doi: 10.1017/S1479262109990232

Nagai, C., Heinig, R., Olano, C. T., Motamayor, J. C., and Schnell, R. J. (2009). "*Fingerprinting of cacao germplasm in Hawaii," Cacao Report No. 1*. Waipahu, HI: Hawaii Agriculture Research Center.

National Institute of Agricultural Biotechnology (2006). *PowerCore (v. 1.0). A program applying the advanced M strategy using heuristic search for establishing core or allele mining sets*. Ranchi: National Institute of Agricultural Biotechnology.

Nieves-Orduña, H. E., Müller, M., Krutovsky, K. V., and Gailing, O. (2021). Geographic patterns of genetic variation among cacao (*Theobroma cacao* L.) populations based on chloroplast markers. *Diversity* 13:249. doi: 10.3390/d13060249

Ochoa, R. (2017). *Jaén y la cultura Marañón*. Lima: La República.

Oddoye, E. O. K., Agyente-Badu, C. K., and Gyedu-Akoto, E. (2013). "Cocoa and its by-products: Identification and utilization," in *Chocolate in Health and*

Nutrition, eds R. R. Watson, V. R. Preedy, and S. Zibadi (Totowa, NJ: Humana Press), 23–37. doi: 10.1007/978-1-61779-803-0_3

Oliva, M., and Maicelo-Quintana, J. L. (2020). Identification and selection of ecotypes of fine native cocoa aroma from the north-eastern zone of Peru. *Rev. Investig. Agroproducc. Sustent.* 4, 31–39. doi: 10.25127/aps.2020 2.556

Oliva-Cruz, S. M. (2020). *Caracterización socioeconómica de la diversidad biológica de cacao Criollo fino de aroma en comunidades rurales de la región Amazonas*. Ph. D. thesis. Chachapoyas-Perú: Universidad Nacional Toribio Rodríguez De Mendoza De Amazonas.

Olivera-Núñez, Q. (2018). *Jaén, Arqueología y Turismo Yanapay Andina Consultores*. Jaén: Municipalidad Provincial de Jaén.

Osorio-Guarín, J. A., Berdugo-Cely, J., Coronado, R. A., Zapata, Y. P., Quintero, C., Gallego-Sánchez, G., et al. (2017). Colombia a source of cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. *Front. Plant Sci.* 8:1994. doi: 10.3389/fpls.2017.01994

Padi, F. K., Ofori, A., Takrama, J., Djan, E., Opoku, S. Y., Dadzie, A. M., et al. (2015). The impact of SNP fingerprinting and parentage analysis on the effectiveness of variety recommendations in cacao. *Tree Genet. Genomes* 11:44. doi: 10.1007/s11295-015-0875-9

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1111/j.1471-8286

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Perrier, X., Flori, A., and Bonnot, F. (2003). "Data analysis methods," in *Genetic diversity of cultivated tropical plants*, eds P. Hamon, M. Seguin, X. Perrier, and J. C. Glaszmann (Montpellier: Enfield Science Publishers), 43–76.

Perrier, X., and Jacquemoud-Collet, J. P. (2006). *DARwin software.* Available online at: http://darwin.cirad.fr/darwin (accessed July 11, 2014).

Phillips-Mora, W., Arciniegas-Leal, A., Mata-Quirós, A., and Motamayor-Arias, J. C. (2013). *Catalogue of Cacao Clones: Selected by CATIE for Commercial Plantings*. Turrialba: CATIE.

Pound, F. J. (1938). "Cacao and witches' broom disease (*Marasmius perniciosus*) of South America," in *Archives Cacao Research*, Vol. 1, ed. H. Toxopeus (Washington DC: American Cacao Research Institute and Brussels), 20–72. doi:

Pound, F. J. (1943). "*Cacao and witches' broom disease (Marasmius perniciosa*)," *Report on a recent visit to the Amazon territory of Peru, September 1942–February 1943*. Trinidad: Yuille's Printery.

Pridmore, R., Crouzillat, D., Walker, C., Foley, S., Zink, R., Zwahlen, M. C., et al. (2000). Genomics, molecular genetics and the food industry. *J. Biotechnol.* 78, 251–258. doi: 10.1016/s0168-1656100000202-00209

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

Rice, R. A., and Greenberg, R. (2000). Cacao cultivation and the conservation of biological diversity. *Ambio* 29, 167–173. doi: 10.1579/0044-7447-29. 3.167

Saavedra-Arbildo, R. P., Cárdenas-Salazar, H., Márquez-Dávila, K. J., Beraun-Cruz, Y., Carranza-Cruz, M. S., Hurtado-Gonzalez, O. P., et al. (2018). "Colecta y estudio de las características morfológicas y organolépticas en fruta fresca y licor de arboles de cacao (*Theobroma cacao* L.) con atributos de poseer caracteristicas de fino y de aroma," in *Proceedings of the International Symposium on Cocoa Research* (ISCR), Lima.

Singh, B. D., and Singh, A. K. (2015). "Mapping populations," in *Marker-Assisted Plant Breeding: Principles and Practices*, eds B. D. Singh and A. K. Singh (New Delhi: Springer), doi: 10.1007/978-81-322-2316-0_5

Tacer-Caba, Z. (2019). "The concept of superfoods in diet," in *The Role of Alternative and Innovative Food Ingredients and Products in Consumer Wellness*, ed. C. M. Galanakis (Amsterdam: Academic Press), doi: 10.1016/B978-0-12-816453-2.00003-6

Thiers, B. (2016). *Index Herbariorum. A global directory of public herbaria and associated staff*. Bronx, NY: New York Botanical Garden's Virtual Herbarium.

Thomas, E., van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., and van Etten, J. (2012). Present spatial diversity patterns of *Theobroma cacao* L. in the Neotropics reflect genetic differentiation in Pleistocene refugia followed by human-influenced dispersal. *PLoS One* 7:e47676. doi: 10.1371/journal.pone.0047676

Torres-Armas, E. A., and Gonzáles-Castro, J. B. (2018). Caracterización de productores en la cadena de valor del cacao fino de aroma de Amazonas. *Conocimiento para Desarrollo* 9, 113–120. doi:

Toxopeus, H. (1985). "Botany, types and populations," in *Cocoa*, 4th Edn, eds G. A. R. Wood and R. A. Lass (London: Longman), 11–37. doi:

Trognitz, B., Scheldeman, X., Hansel-Hohl, K., Kuant, A., Grebe, H., and Hermann, M. (2011). Genetic population structure of cacao plantings within a young production area in Nicaragua. *PLoS One* 6:e16056. doi: 10.1371/journal. pone.0016056

Valdez, F. (2013). "Prefacio," in *Arqueologia Amazonica: las civilizaciones ocultas del bosque tropical*, ed. F. Valdez (Quito: IRD Editions). doi:

Waits, L. P., Luikart, G., and Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Mol. Ecol.* 10, 249–256. doi: 10.1046/j.1365-294X.2001.01185.x

Wang, B., Motilal, L. A., Meinhardt, L. W., Yin, J., and Zhang, D. (2020). Molecular characterization of a cacao germplasm collection maintained in Yunnan, China using single nucleotide polymorphism (SNP) markers. *Trop. Plant Biol.* 13, 359–370. doi: 10.1007/s12042-020-09267-y

Wickramasuriya, A., and Dunwell, J. M. (2018). Cacao biotechnology: current status and future prospects. *Plant Biotechnol. J.* 16, 4–17. doi: 10.1111/pbi.12848

Xu, W. (2016). *Functional Nucleic Acids Detection in Food Safety: Theories and Applications*. Singapore: Springer, doi: 10.1007/978-981-10-1618-9

Yacenia Morillo, C., Morillo, A. C., Muñoz, F. J. E., Ballesteros, P. W., and González, A. (2014). Caracterización molecular con microsatélites amplificados al azar (RAMs) de 93 genotipos de cacao (*Theobroma cacao* L.). *Agronomia Colombiana* 32, 315–325. doi: 10.15446/agron.colomb.v32n3.46879

Zarrillo, S., Gaikwad, N., Lanaud, C., Powis, T., Viot, C., Lesur, I., et al. (2018). The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. *Nat. Ecol. Evol.* 2, 1879–1888. doi: 10.1038/s41559-018-0697-x

Zhang, D., Arevalo-Gardini, E., Mischke, S., Zuñiga-Cernandes, L., Barreto-Chavez, A., Adriazola, et al. (2006a). Genetic diversity and structure of managed and semi-natural populations of cacao (*Theobroma cacao*) in the Huallaga and Ucayali valleys of Peru. *Ann. Bot.* 98, 647–655. doi: 10.1093/aob/mcl146

Zhang, D., Mischke, S., Goenaga, R., Hemeida, A. A., and Saunders, J. A. (2006b). Accuracy and reliability of high-throughput microsatellite genotyping for cacao clone identification. *Crop Sci.* 46, 2084–2092. doi: 10.2135/cropsci2006.01. 0004

Zhang, D., Gardini, E. A., Motilal, L. A., Baligar, V., Bailey, B., Zuñiga-Cernades, L., et al. (2011). Dissecting genetic structure in farmer selections of *Theobroma cacao* in the Peruvian Amazon: implications for on farm conservation and rehabilitation. *Trop. Plant Biol.* 4, 106–116. doi: 10.1007/s12042-010-9064-z

Zhang, D., Martínez, W. J., Johnson, E. S., Somarriba, E., Phillips-Mora, W., Astorga, C., et al. (2012). Genetic diversity and spatial structure in a new distinct *Theobroma cacao* L. population in Bolivia. *Genet. Resour. Crop Evol.* 59, 239–252. doi: 10.1007/s10722-011-9680-y

Zhang, D., and Motilal, L. (2016). "Origin, dispersal, and current global distribution of cacao genetic diversity," in *Cacao Diseases*, eds B. A. Bailey and L. W. Meinhardt (Switzerland: Springer International Publishing), 1–33.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 11 October 2022
DOI 10.3389/fgene.2022.1023870

Multiple migrations from East Asia led to linguistic transformation in NorthEast India and mainland Southeast Asia

Debashree Tagore[1], Partha P. Majumder[1,2], Anupam Chatterjee[3,4] and Analabha Basu[1]*

[1]National Institute of Biomedical Genomics, Kalyani, India, [2]Indian Statistical Institute, Kolkata, India, [3]Department of Biotechnology, North-Eastern Hill University, Shillong, India, [4]School of Biosciences, Royal Global University, Guwahati, India

NorthEast India, with its unique geographic location in the midst of the Himalayas and Bay of Bengal, has served as a passage for the movement of modern humans across the Indian subcontinent and East/Southeast Asia. In this study we look into the population genetics of a unique population called the Khasi, speaking a language (also known as the Khasi language) belonging to the Austroasiatic language family and residing amidst the Tibeto-Burman speakers as an isolated population. The Khasi language belongs to one of the three major broad classifications or phyla of the Austroasiatic language and the speakers of the three sub-groups are separated from each other by large geographical distances. The Khasi speakers are separated from their nearest Austroasiatic language-speaking sub-groups: the "Mundari" sub-family from East and peninsular India and the "Mon-Khmers" in Mainland Southeast Asia. We found the Khasi population to be genetically distinct from other Austroasiatic speakers, i.e. Mundaris and Mon-Khmers, but relatively similar to the geographically proximal Tibeto Burmans. The possible reasons for this genetic-linguistic discordance lie in the admixture history of different migration events that originated from East Asia and proceeded possibly towards Southeast Asia. We found at least two distinct migration events from East Asia. While the ancestors of today's Tibeto-Burman speakers were affected by both, the ancestors of Khasis were insulated from the second migration event. Correlating the linguistic similarity of Tibeto-Burman and Sino-Tibetan languages of today's East Asians, we infer that the second wave of migration resulted in a linguistic transition while the Khasis could preserve their linguistic identity.

KEYWORDS

Austroasiatic, Khasi, Tibeto Burman, admixture, migration, linguistic transformation

OPEN ACCESS

EDITED BY
Sibte Hadi,
Naif Arab University for Security Sciences (NAUSS), Saudi Arabia

REVIEWED BY
Hui Li,
Fudan University, China
Randy J. LaPolla 羅仁地,
Nanyang Technological University, Singapore

*CORRESPONDENCE
Analabha Basu,
ab1@nibmg.ac.in

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

RECEIVED 20 August 2022
ACCEPTED 27 September 2022
PUBLISHED 11 October 2022

CITATION
Tagore D, Majumder PP, Chatterjee A and Basu A (2022), Multiple migrations from East Asia led to linguistic transformation in NorthEast India and mainland Southeast Asia.
Front. Genet. 13:1023870.
doi: 10.3389/fgene.2022.1023870

Introduction

The Indian subcontinent is genetically one of the most diverse regions of the world harboring over 1.25 billion people (2011 census). The region has been part of the earliest waves of Anatomically Modern Human (AMH) migrations which peopled South and Southeast Asia, including Australia, beginning around 60,000 years ago (Basu et al., 2003; Kivisild et al., 2003; Macaulay et al., 2005; Thangaraj et al., 2005). Over time, the region has also witnessed multiple waves of migration (Basu et al., 2003; Endicott, Metspalu, and Kivisild 2007; Majumder 2008; Basu et al., 2016) that has contributed to its huge genetic, linguistic, and cultural diversity. The Indian subcontinent is bounded in the North and Northeast by the Himalayas. NorthEast India (NEI) is a unique region that is bordered in the north by the high ranges of eastern Himalayas and two-thirds of it is intermediate hilly terrain, interspersed by fertile riverbeds and flat valleys. The population density also varies accordingly; while the river valleys are densely populated and cosmopolitan, the highlands are sparsely populated by small isolated ethno-lingual groups. Major population groups that reside here speak Tibeto-Burman languages, which belong to the non-Sinitic phylum of the Sino-Tibetan language family. Now restricted by political boundaries, this region is likely to have been a land bridge between peninsular India (PI) and Mainland Southeast Asia (MSEA) and has been an active corridor of migration and admixture of different ethnolinguistic populations in the past (Gadgil et al., 1993; Cavalli-Sforza et al., 1994; Reddy et al., 2007; Tagore et al., 2021; Liu et al., 2022) and hence should be considered in continuum with the population demographic history of East and Southeast Asia. Individuals from mainly five language families reside in NEI and the neighborhood of MSEA: namely Sino-Tibetan, Tai-Kadai, Hmong-Mein, Austronesian, and Austroasiatic (AA). However, more recent migrations of the ancestors of Indo-European language speakers of India, who possibly entered India through the northwestern corridor also had a large impact on the populations of NEI (Gayden et al., 2009; Basu et al., 2016). The Austroasiatic language family comprises three major subfamilies: Munda, Mon-Khmer, and Khasi-Khmuic (Diffloth Gerard 2005a). Within the Austroasiatic family, the Khasi language (the sole language of the Khasi-Khmuic branch of the Austroasiatic language family in India) is spoken in NEI mainly in parts of the north-eastern state of Meghalaya.

In this study we look into the population genetics of the Khasi, residing amidst the Tibeto-Burman speakers as an isolated population. These Khasi speakers are separated by large physical distance from their nearest Austroasiatic language-speaking sub-groups: the Munda sub-family from East and peninsular India and the Mon-Khmer sub-family in Mainland Southeast Asia. Here, we dissect the genetic relationship of the Khasis with the other Austroasiatic subgroups and in an attempt to do so, reconstruct the population history of NorthEast India, and the neighboring East and Southeast Asia, in the context of the Khasi Austroasiatics.

Despite a strategic location, most genetic studies on NEI populations have been done using either uniparental markers (Cordaux et al., 2003; Borkar et al., 2011) or a small number of autosomal markers (Maity, Nunga, and Kashyap 2003; Krithika et al., 2005; Krithika et al., 2006; Mastana et al., 2007; Gayden et al., 2009). Cordaux et al. (2003) mitochondrial DNA (mtDNA) and Y chromosome-based study suggest two possibilities regarding the peopling of NEI: either TBs were the earliest inhabitants, or the TB replaced the Austroasiatic (AA) inhabitants of NEI. Using microsatellite data, and comparing the Khasi-Khmuic speakers with their neighboring Tibeto-Burman speakers, showed the populations to be extremely homogeneous (Langstieh et al., 2004) a fact further supported by later studies with mtDNA and Y-chromosome (Cordaux et al., 2004). Initially, researchers came up with opposing views on the origin of TB populations. One theory, based on Y-chromosome analyses, suggests that the TB ancestors originated in the upper and middle Yellow River basin (Su et al., 2000). Another theory suggests the Yangtze River as their ancestral source followed by the northward movement to the Yellow River basin (Van Driem 2005). In our previous study (Tagore et al., 2021) we also proposed a theory where we suggested that present-day Tibeto Burmans were likely Austroasiatics in the past, who were part of the earliest settlers of the region (Hill et al., 2006). Y-chromosome based study by Wang et al. (2018) suggested that the peopling of the Tibetan plateau by Tibeto Burman ancestors happened some 40KYA (40 thousand years ago). This coincides with the presence of hunter gatherers in this region. However, it was during the Neolithic period, ~6KYA, when the expansion of different Y chromosome lineages was observed leading to the present day distribution of the TBs. This time coincides with the migration of East Asians in MSEA (Tagore et al., 2021). Yu et al. (2021) have suggested that migration of both Yellow river basin millet farmers and Yangtze river basin rice farmers contributed to different linguistic and genetic groups in MSEA. They have also proposed that around 6KYA, people from the middle Yellow River Basin migrated south-westward and mixed with the local population to give rise to the initial TBs. Basu et al. (2003) has shown that the TB and AA speakers of India are similar in their mtDNA profile but harbor very distinct Y-chromosomes. Our previous study (Tagore et al., 2021) observed the genetic relatedness between the Tibeto Burmans and Austroasiatic speakers (Mon Khmers) of Malaysia, because of ancient shared ancestry as well as owing to gene flow in both these populations from East Asia. Another study (Guo et al., 2022) found the present day TBs to cluster between the millet cultivators of Yellow River basin as well as the Austroasiatic speakers of Southeast Asia in a Principal Components Analysis. In further analyses they found southern Tibeto Burmans were genetically closest to the AAs.

The Northeast Indian populations were clustered with populations of East and Southeast Asia than with mainland Indians (Langstieh et al., 2004; Basu et al., 2016; Tagore et al., 2021). Other studies on the mtDNA hypervariable region and autosomal microsatellite markers found that despite the present political boundary, the Tibeto-Burman speakers from NEI showed a closer genetic affinity with East Asian populations than with other mainland Indian populations (Cordaux et al., 2003; Krithika, Maji, and Vasulu 2008; Basu et al., 2016). This further supports the fact that ancient migration events occurred through the NEI corridor before the political boundaries were drawn (Basu et al., 2003; Basu et al., 2016). Such genetic studies are in agreement with the linguistics of Northeast India: the Tibeto Burman language group is closely related to the languages of East Asia. Apart from the genetic similarity with the East Asians, the TB also shows some complex admixture with other Indian populations belonging to different ancestries. The TBs harbor genetic ancestry predominantly in Indo-European speakers (henceforth referred to as ANI or Ancestral North Indian) who mainly reside in the northern part of India, and also genetic ancestry predominantly in Dravidian language speakers (henceforth referred to as ASI or Ancestral South Indian) who are almost exclusively confined to the southern part of India (Basu, Sarkar-Roy, and Majumder 2016).

The Khasis are one of the few populations in the world that follow a matrilineal system of inheritance. Besides the linguistic similarity, anthropologists and archaeologists have established that the Khasis have cultural similarities with Mundaris and Mon-Khmer populations. It has been shown that they share similar stone tools and have similar death rituals of erecting memorial stones for the deceased (Gurdon 1914). Linguistically, the Khasi language is more similar to languages of the Mon-Khmer branch than those of the Mundari branch and linguists have often assigned Khasi and Mon-Khmer languages to the same group (Pinnow et al., 1942; Chazée 1999). Khasi language also bears lexical and morphological similarities to some Tibeto Burman languages (Longmailai 2015). Peiros suggested a significant number of words were similar between Proto Austroasiatics and Proto-Sino-Tibetans (Peiros 2011).

Nevertheless, the presence of ancient Austroasiatics (AA) speakers across NEI still remains a possibility. Our previous study (Tagore et al., 2021) on autosomal data of the Mundari and Mon-Khmer Austroasiatics indicated that in pre-Neolithic times, the ancestors of today's Austroasiatic speakers had a widespread distribution possibly extending from Central India to Southeast Asia (SEA), further supported by Lipson et al. (2018). They were later in time fragmented and isolated to small pockets resulting in their present-day disjoint geographic distribution. What is intriguing is that given the widespread distribution of Austroasiatic speakers from Central India to SEA across NEI and the central location of the Khasis, it is possible that the Khasis will serve as a genetic link between the two Austroasiatic groups on either side of NEI.

There have been very few studies on the genetics of Khasi, so as to reach any plausible conclusions. One previous study on uniparental markers has proposed a genetic continuity between the Mundari Austroasiatics of Central India, Khasi-Khmuic, and Mon-Khmer (Reddy et al., 2007). Using multidimensional scaling of the pairwise F_{ST} distances calculated on Y-haplogroups of Austroasiatics and neighboring populations, they found the three Austroasiatic groups (Mundari, Khasi-Khmuic, and Mon-Khmer) to cluster together. They also found the Y haplogroup O-M95, restricted within the Austroasiatics and postulated to have originated in the Mundaris, is present in the Khasis at a frequency intermediate to that of Mundaris and Mon-Khmers. They suggested an initial presence of Austroasiatics in Central India with rapid migration to Southeast Asia *via* the Northeast corridor carrying the O-M95 haplogroup.

The cultural and linguistic similarities of the Khasis with other Austraoasiatic groups as mentioned earlier prompt us to investigate their genetic affinities. The geographic location of the Khasis also makes it imperative to investigate the impact of East Asian migrations on the genetic make-up of the Khasis.

Materials and methods

Dataset preparation and quality control

DNA samples from 22 individuals speaking the Khasi language were sequenced and merged with the genotype dataset of 1,451 individuals that were used in our previous study (Tagore et al., 2021) using PLINK(Shaun et al., 2007). The details of the datasets are provided in Supplementary Table S1A–C. Only biallelic loci were included in our analysis. We removed all monomorphic variants and SNPs with alleles A/T and G/C from our analysis. We also removed SNPs with missingness of more than 5% in the entire dataset, or SNPs that were missing in more than 25% of individuals in any of the 15 subpopulations (second column of Supplementary Table S1A). We also excluded SNPs that were out of Hardy Weinberg equilibrium ($p < 10^{-6}$) in any of the 15 subpopulations. This combined dataset had 310110 SNPs.

Principal components analysis

In order to understand the overall population structure and the genetic affinities of the individuals in our dataset, we performed Principal Components Analyses (PCAs) using the smartpca program of the EIGENSOFT package (Patterson, Price, and Reich 2006). We performed an initial PCA on all the mainland Indians (all Indian populations excluding those belonging to the "Island" group as in Supplementary Table

S1A) and Malaysian populations. We considered the first two Principal Components (PCs) to visualize the data.

A second PCA was run on a subset of populations used in the first PCA. This subset was chosen based on linguistic similarity and geographic proximity to the Khasis. Thus, we included the Austroasiatics from Central India (AACI), Austroasiatics of Malaysia (AAM), Khasi, and Tibeto Burmans (TBs).

TreeMix

In order to understand how populations were related to each other through a common ancestor and the impact of genetic drift, we built ancestry graphs using TreeMix (Pickrell et al., 2012) version 1.12. Such graphs were created with AACI, AAM, TB, and the Khasi populations using the Mbuti Pygmies from Africa as an outgroup.

F_{st} estimates

Using PLINK (Shaun et al., 2007) version 1.9, the weighted F_{st} between each subpopulation of AACI, AAM, TB, and the Khasi was estimated. These values were rounded to the third decimal place.

Outgroup *f3* statistics

Outgroup *f3* statistics measures the shared drift between two populations relative to an extremely diverged population outgroup. Using ADMIXTOOLS (v5.1) (Patterson et al., 2006), we calculated outgroup *f3* statistics of the form *f3* (Mbuti Pygmy; Khasi, Y) where Mbuti Pygmy was the outgroup. Y was AACI, AAM, and TB subpopulations.

ADMIXTURE analysis

To infer the different ancestral components present in the admixed populations and the proportions of each such component in an individual's genome, we performed unsupervised clustering as implemented in ADMIXTURE (Alexander, Novembre, and Lange 2009) (v1.3.0). We ran ADMIXTURE using all Indian populations (AACI, ANI, ASI, ATB, and Khasi), Malaysian populations (AAM and ANS), and all East Asians. We ran ADMIXTURE by sequentially increasing the number of clusters, which corresponds to the number of identified ancestries (k), in each run of the analysis on a given dataset. ADMIXTURE estimates the proportion of each of the k ancestries in the genome of each individual of the dataset and also computes a cross-validation error (CVE) for that particular run. Standard error was estimated for the ancestry proportion estimates at the minimum CVE using the moving block bootstrap approach as implemented in ADMIXTURE.

Admixed segment length calculation

Dataset was phased using SHAPEIT v2.r790 (Delaneau, Marchini, and Zagury 2012). From the phased dataset we extracted the phased genomes for Khasi, Jamatia, Miazou (a Southern East Asian-like ancestry population), Yakut (a Northern East Asian-like ancestry population), and Kshatriya (an ANI-like ancestry population). This was followed by local ancestry estimation using RFMix (Maples et al., 2013) version 1.5.4, to identify regions of genomes of Khasi and a TB population (in this case Jamatia) corresponding to different ancestries. The different ancestries which we considered for the local ancestry estimation were inferred from the ADMIXTURE run where the CVE was minimum, i.e. at $k = 8$. Ancestries for which tract lengths were estimated in Khasi included: "Southern EA-like", "Jehai-like", "MahMeri-like", "Birhor-like (AACI-like)" and "ANI-like". In addition to these ancestries, tract lengths corresponding to "Northern EA-like" ancestry were also estimated for Jamatia. It is to be noted here that the Northern EA-like ancestry was practically absent in the Khasis. We plotted the cumulative distribution of these tract lengths to compare the sizes of these tract lengths corresponding to the different ancestries in both Khasi and Jamatia.

Estimating admixture time

Gene flow events between genetically distinct populations create linkage disequilibrium between all loci that are highly differentiated between the two ancestral populations. Segments resulting from admixture follow an exponential distribution, where as a result of recombination, this linkage disequilibrium pattern declines exponentially over time and from which the number of generations since admixture can be estimated (Racimo et al., 2015). To date the 'time since the last admixture event' between different populations, we generated "co-ancestry curves" using MOSAIC (Salter-Townshend and Myers 2019) (v1.2). Here the closest surrogate populations were chosen as "donors". Coancestry curves measure how often, in an admixed ("recipient") population; a pair of haplotypes has been inherited from each respective donor population. Given a single admixture event, ancestry chunks inherited from each source, reduce in size because of recombination, resulting in an exponential decay of these coancestry curves. The time (in generations) since admixture is calculated from the rate of decay in the curves.

To detect 2-way admixture events in Jamatia and Khasi, we used Yakut as a surrogate donor of the "Northern EA-like"

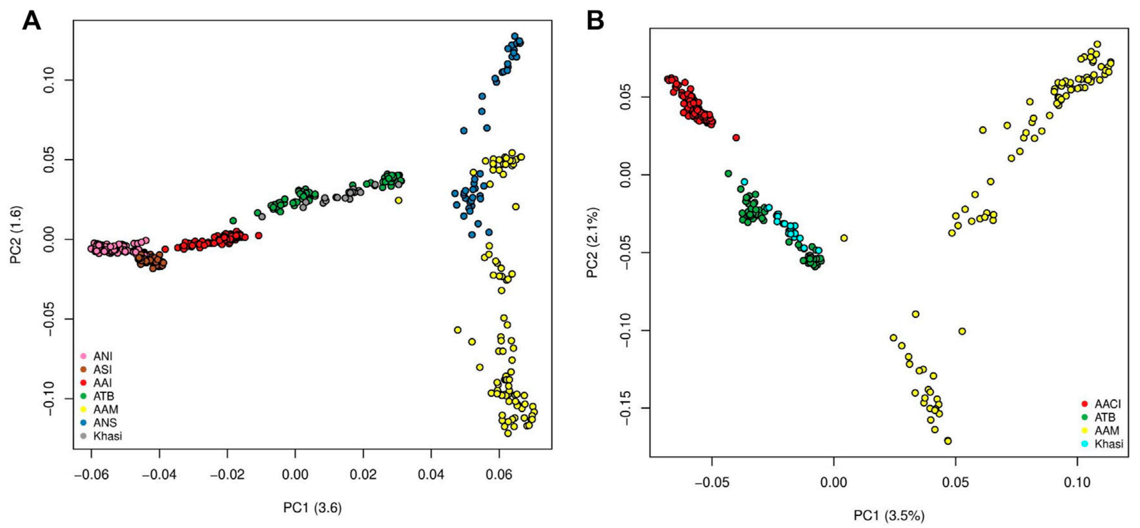


FIGURE 1
(A) PCA on Mainland Indian and Malaysian populations (ANI: Ancestral North Indian, ASI: Ancestral South Indian, AACI: Austroasiatics of Central India, ATB: Ancestral Tibeto Burmans, AAM: Austroasiatics of Malaysia, ANS: Austronesians); **(B)** PCA on the Austroasiatics and the Tibeto Burmans.

ancestry, Miazou for “Southern EA-like” ancestry and Birhor for “Austroasiatic-India or Mundari” ancestry. We chose Khasi and Jamatia as recipients (having ancestry from each of the source populations as a result of admixture) and estimated the time since the last admixture between the donors. We created co-ancestry curves for the surrogate donors (Miazou and Birhor) in both Khasi and Jamatia and another co-ancestry curve for donors Yakut and Birhor in Jamatia. The rate of decay in the curves was calculated which was equal to the number of generations since admixture took place.

Results

The first two Principal Components (PCs) of the PCA with all the mainland Indians (all Indian populations excluding those belonging to the “Island” group as in Supplementary Table S1A) and Malaysian populations explained 3.6% and 1.6% of the variation. In PC1-PC2 space the individuals belonging to the major population groups (as classified in the second column of Supplementary Table S1A), formed unique clusters. In the PC1 axis the Indian population, specifically the Ancestral North India-like (ANI-like) populations were on one extreme while the Malaysian populations were on the other. While most Indian populations were distinguishable along the first PC, the two Malaysian populations separated along the second PC. We found the Khasis to cluster with the Tibeto Burmans (Figure 1A) (It is to be noted that this is very similar to Figure 1D in Tagore et al., 2021 where we had all the populations except the Khasis). Using a smaller subset of the above we did a second PCA, where, we considered the Khasis along with the two Austroasiatic groups from our previous study i.e. Mon-Khmer speaking Austroasiatics from Malaysia (AAM), Mundari speaking Austroasiatics from Central India (AACI). We also included the Tibeto-Burman population (TB) who were geographically proximal to the Khasis (Figure 1B; Supplementary Figure S1). We considered three Principal Components (PCs) which together could explain 7.1% of the total variation. In the PC1-PC2 space, we found that the three Austroasiatic groups formed distinct clusters. The Khasi did not cluster with either of the other two Austroasiatic populations, instead, clustered with the TB subgroups (Figure 1B). Khasi and TB were distinguished as separate clusters in PC3. Here we could also identify two separate clusters within the TB: one comprising Jamatia and Tripuri (that clustered closer to the AAM) and the other comprising M-Brahmin and Tharu (Supplementary Figure S1), who are known to have been admixed with other populations of North India and the Upper Gangetic plains (Basu, Sarkar-Roy, and Majumder 2016). A similar pattern of clustering was found in the TreeMix analysis (Supplementary Figure S2). The Khasis clustered with the Tibeto Burmans in a branch separate from the other two Austroasiatic populations. The Munda and Mon-Khmer populations also formed distinct clusters.

On the same set of the population (as used in Figure 1B), we surveyed the allele frequencies and calculated pairwise F_{st} (Weir and Cockerham 1984) between them using PLINKv1.9. Here again, we found F_{st} between the Khasi and TB groups to be low

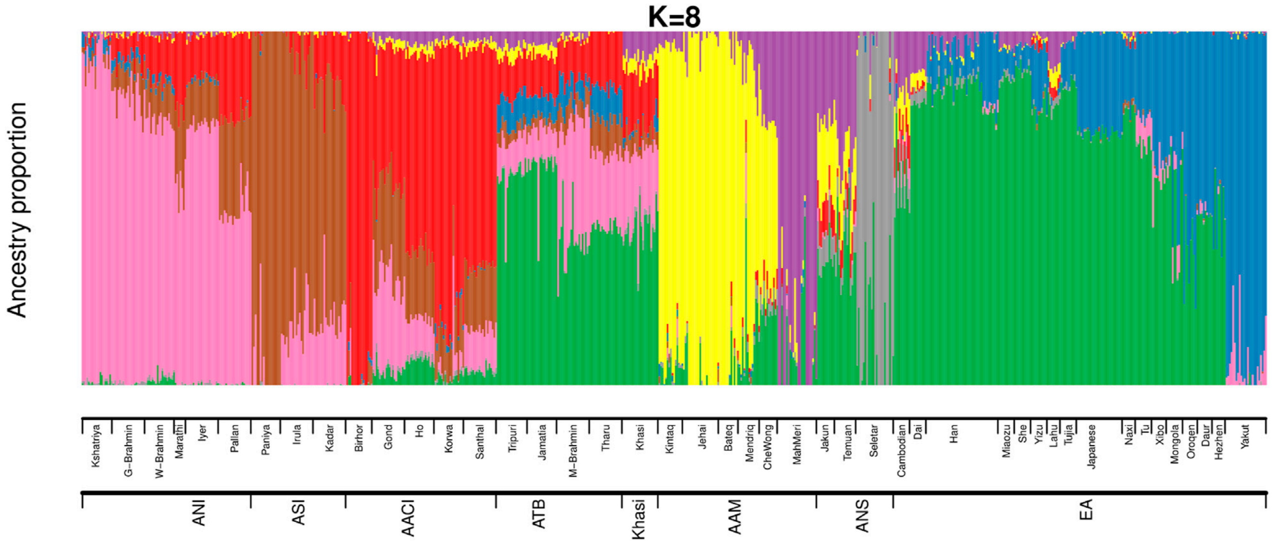


FIGURE 2
ADMIXTURE analysis on Mainland Indian populations, Malaysians and East Asians of HGDP (ANI: Ancestral North Indian, ASI: Ancestral South Indian, AACI: Austroasiatics of Central India, ATB: Ancestral Tibeto Burman, AAM: Austroasiatics of Malaysia, ANS: Austronesians, EA: East Asians of HGDP).

(mean = 0.019) (Supplementary Figure S3) which was even lower than those between Khasi and AAM (mean = 0.061) and between Khasi and AAI (mean = 0.033).

In our analysis, *f3* (Mbuti Pygmy; Khasi, X) we used the African Mbuti Pygmy as the outgroup. We measured the *f3* values of Khasi with AAM, AACI, and TB (details in Materials and Methods, Supplementary Table S3). The mean *f3* values are highest (mean = 0.279) between Khasi and TB, indicative of an exclusive and recent shared genetic history. The mean *f3* values were higher (mean = 0.277) for Khasi-AAM than between Khasi-AAI (mean = 0.265). The apparent discordance in the pattern of f3 and F_{st} values when Khasi are compared to AACI and AAM is largely due to the fact that F_{st} is affected by drift. It is to be noted here that we see the AAM populations be highly drifted in our Treemix analyses (Supplementary Figure S2).

We then estimated the genomic ancestries and admixture proportions at an individual level by considering all populations from India and Malaysia (the same populations we used in Figure 1A). We also included the East Asians in this analysis because we had observed in our previous study (Tagore et al., 2021) that an East Asian ancestry component was found among some Austroasiatic populations. We did an ADMIXTURE (Alexander, Novembre, and Lange 2009) analysis (Figure 2; Supplementary Figures S4A,B) where the Cross-Validation Error (CVE, details in Materials and Methods) was minimized at $k = 8$ (Supplementary Figure S4A). We also calculated the ancestry proportions for each of these populations (Supplementary Table S3).

The ancestry proportions of the Khasi, as estimated by ADMIXTURE, are distinct from the other two Austroasiatic groups (AAM and AACI) but are very similar to that of the Tibeto Burmans, especially to the Jamatia and Tripuri. The major ancestry identified among the Khasi was also the predominant ancestry identified among the Southern-EA populations (like Dai). In the ADMIXTURE plot (Figure 2), it is depicted by the green color (nearly 44%, "Southern-EA-major" in Supplementary Table S3). This green-colored component is the ancestry modal to the Southern East Asians (henceforth referred to as "Southern EA-like" ancestry) such as Dai. The Khasi genome also has a substantial proportion of AACI-like ancestry (16%, "AACI-major"; red in color) and AAM-like ancestry (4%; yellow color modal to Jehai and 7%; purple color modal to MahMeri). 20% of the Khasi genome is of ANI-like ancestry ("ANI-major", pink in color). Neither Khasi nor the TB groups had in them any distinctly identified "Khasi-like" or "TB-like" component respectively. Alternatively, the AACI and AAM had genomic components mostly exclusive to them: 64% "AACI-major" component (red in color) and 62% "AAM-major" components (yellow in color) respectively. The East Asian component present in Khasis (green in color) was also present in AACI and AAM, though in lesser proportions of 3.5% and 7.6% respectively. The TBs, on the other hand, had an even higher proportion of this component (51%). In addition to this East Asian component, TBs also have a substantial proportion (7.4%) of a second East Asian component (blue in color). This East Asian component (henceforth referred to as "Northern-EA-major") is modal in the East Asian populations residing in today's Northern China (e.g. the Yakut). It is to be noted here that these "Northern-EA" and "Southern-EA"

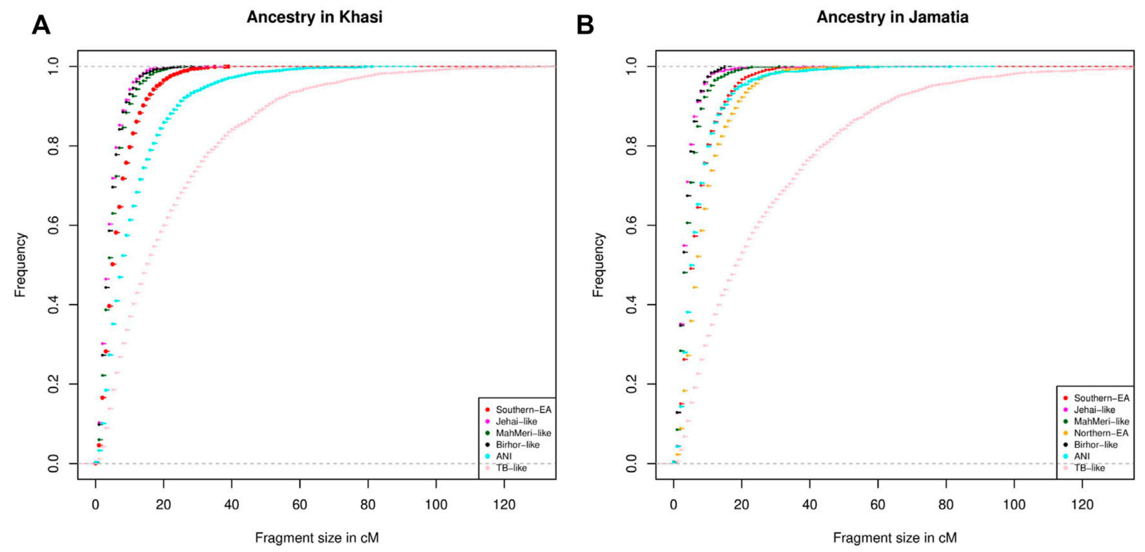


FIGURE 3
(A) Distribution of tract lengths corresponding to different ancestries in Khasi; **(B)** Distribution of tract lengths corresponding to different ancestries in Jamatia.

components were also identified in our previous study. Although the TB (particularly Jamatia and Tripuri) and Khasi individuals cluster close in the PCA, this ancestry is negligible in the Khasis. Compared to the Southern EA-like ancestry, the Northern EA-like ancestry is negligible in the other two Austroasiatics groups (AACI and AAM) as well.

To investigate the chronology of admixture events into Khasi, we performed local ancestry estimation, using RFMix (Maples et al., 2013). We identified regions within genomes of Khasi individuals representing different ancestries as inferred in the ADMIXTURE analysis. We estimated the length of admixed tracts representing the following ancestries: "Jehai-like", Birhor-like", "MahMeri-like", ANI-like" and "Southern -EA-like". We looked into the cumulative frequency distribution of these tract lengths. Larger tracts (segments) would correspond to a recent introduction of the corresponding ancestry, and hence can be used as an indicator of the sequence of admixture events. We found that the tract lengths corresponding to Birhor-like and Jehai-like ancestry are the smallest in the Khasis. This is followed by MahMeri-like, ANI-like and Southern EA-like ancestry tracts. This indicates that Southern EA-like ancestry is most recently introduced in the Khasis (Figure 3A).

We then repeated the same analysis with the Jamatia (a subgroup of the TB) and with the same five ancestries i.e. Birhor-like, Jehai-like, MahMeri-like, ANI-like, and Southern EA-like ancestries. Furthermore another ancestry: the Northern EA-like was also included in the analyses because this was an additional ancestry present substantially in the TBs (as evident from ADMIXTURE analysis) but was absent among the Khasis. Similar to what was observed in the Khasis, tract lengths corresponding to the Birhor-like and Jehai-like ancestry are the smallest followed by MahMeri-like, ANI-like, Southern EA-like ancestries. This mimics the chronology of the admixture events that we see in the Khasis. However, the largest length of admixture tracts corresponded to the Northern EA-like ancestry (Figure 3B). This indicates that though the overall sequence of admixture i.e. introduction of ancestries within the Khasis and TBs are similar, the introduction of the Northern EA-like ancestry is the most recent event and unique to the TBs. Thus we conclude that the admixture of the East Asian populations and ancestors of present-day Khasi and Tibeto Burmans is a relatively recent event; of the two distinct East Asian genetic ancestries, the Northern-EA ancestry was introduced in the Tibeto Burmans subsequent to the Southern EA-like ancestry. While both Khasis and TBs have experienced multiple admixture events, the Northern East Asian admixture largely with the TBs is the one which is unique and recent.

We further dated these local admixture events using a method implemented in MOSAIC (Salter-Townshend and Myers 2019) that infers admixture time by fitting an exponential decay coancestry curve (details in Supplementary Material). We chose homogeneous representative populations such as Yakut for Northern EA-like ancestry, Miazou for Southern EA-like ancestry and Birhor for Central Indian Austroasiatic ancestry, as a source population for admixture in populations such as Khasi and Tibeto-Burman. We found that the last evidence of admixture between Southern EA-like ancestry-bearing populations and Austroasiatics and TB took place 13.9 and 10.5 generations ago when Khasi and Jamatia (a representative subgroup for the Tibeto Burman population) were chosen as recipients (Supplementary Figures S5A,B). We also found that the incorporation of Northern EA-like ancestry in the Jamatia happened as recently as 8.3 generations ago (Supplementary Figure S5C). These findings were in accordance with the chronology of events we inferred from the RFMix analysis. This substantiates our conclusion that there were at least two distinct admixture events in NEI populations with East Asians where populations bearing Southern EA-like ancestry admixed first with both the ancestors of TB and Khasi and later populations bearing Northern EA-like ancestry admix mostly with the ancestors of TB.

Discussion

The Khasi are a relatively large population, subdivided into groups owing to geographical barriers, and show considerable heterogeneity as evident from an anthropometric study (Das 1970). We find from our analyses (PCA, ADMIXTURE, TreeMix), that the Khasis are genetically very similar to the Tibeto-Burmans. The PCA cannot identify the Khasis as a distinct cluster, separate from other TB populations when compared with AACI and AAM. The Khasi Austroasiatics are distinct from the other Austroasiatics (Mundari or AACI and Mon-Khmer or AAM) in our study which conforms to the linguistic classification by Diffloth (2005b). The ancestral components inferred using ADMIXTURE in the Khasis and TBs are also very similar. In our previous study, we had observed that the AACI, TB, and AAM shared a deep common ancestry and proposed that all of their ancestors likely spoke some proto-AA language. The observed genomic profile of the Khasis suggests that the Austroasiatic-speaking Khasis fit well into the proposed model. We had also postulated that the ancestors of the present-day TB and the AAM populations experienced admixture with southward migrating EA agriculturists. Here we find that the Khasis, residing in the same region as the TBs, experienced the same sequence of admixture events as the Jamatia (TB). This indicates they likely share a common history. The southward migration of East Asians led to the incorporation of East-Asian ancestry in the Tibeto Burmans and the Khasis. This migration was extensive and as we have also previously observed, the admixture signals of this migration can also be found among other AA speakers (predominantly the AAM).

Despite an overall genomic similarity of the TBs and Khasis, there is a distinct difference between the TB populations and the Khasis: unlike the TBs, the Khasis lack Northern East Asian ancestry. Results from our RFMix analysis suggest that there were at least two distinct waves of East Asian migration. The first wave brought the Southern East Asian ancestry that got incorporated in both the Khasis and the TBs and the second wave brought the Northern East Asian component. This migration started possibly from Northern EA and led to the introduction of the Northern EA-like genomic ancestry into the TB population but not in the Khasis. It is to be noted that the Northern EA component is absent from other Austroasiatic speakers as well (AACI and AAM), although some of these populations have substantial EA ancestry, i.e Southern EA ancestry. Wang et al. (2018) suggested that the TBs were an admixed group resulting from two distinct ancient populations: a hunter-gatherer population, (which we believe were the proto-Austroasiatics) and a millet farmer population from middle Yellow River basin. A genetic link between the millet farming proto-Sino Tibetans of the Yellow River basin and Tibeto Burmans has also been proposed by Guo et al. (2022). Though Wang et al propose a two wave migration leading to the formation of TBs, they propose that out of the two, only one wave of migration formed both the TBs of India and populations of MSEA. However, our study suggests that though there were atleast two waves of migration, the second wave solely affected the TBs while the first affected both the TBs and the AAs of MSEA.

We, therefore, propose, in agreement with our previous study, that the ancestors of extant Austroasiatic speakers were widespread across Central India and Southeast Asia encompassing the present-day location of the TBs and Khasis. This is in agreement with other studies (Cordaux et al., 2004). It is hence plausible that the ancestors of present-day Tibeto-Burman speakers spoke some form of an Austroasiatic or Proto-Austroasiatic language. With time, the Austroasiatic populations evolved into three major branches as we see them today namely Mundari, Mon-Khmer, and Khasi.

Higham has suggested that before Neolithic expansion, this region was inhabited by hunter gatherers (Higham 2017). He also suggested that the expansion of farming communities happened from two regions that reached mainland Southeast Asia: one of millet cultivators from the Yellow River basin and another of the rice cultivators from the Yangtze River basin. Such migration events are also supported by morphological studies. Cranial (Matsmura 2011) and Dental (Matsmura 2010) morphological studies found two groups of individuals at the Man Bac excavation site in Southeast Asia: one close to the Neolithic inhabitants of Weidun in the Yangtze Valley and the other to the local hunter gatherers. Archaeological studies in Southeast Asia also supports presence of hunter-gatherers in Southeast Asia as well as Southern China (Higham 2013) and that archaeological sites provide indication that immigrants from Southern China encountered these hunter gatherers on their way. Infact, Neolithic migration has also diluted the genetic differentiation within China (Yang et al., 2020). An extensive documentation of rice spread also supports the spread of rice from China to Southeast Asia (Fuller et al., 2010).

We argue that the language of the extant TBs is a result of this linguistic shift, possibly evidence of elite dominance, which is a consequence of the migration and gene flow from Northern East Asia. When we look at the ancestral segments of Northern-EA ancestry, we find that they are among the longest ancestral segments in TB, preceded by segments of Southern EA-like ancestry. We postulate that the two migration events from East Asia were such that initially, populations bearing Southern EA-like ancestry arrived in NEI, and later came the populations of Northern EA-like ancestry. The Southern EA-like ancestral segments are also present in Khasis and AAM, the two Austroasiatic groups with substantial East Asian ancestry. In these populations, Southern EA-like ancestral segments are among the longest. The AACI however have negligible East Asian components in their genome. It is to be noted here that the language of the AACI, i.e. Mundari is much more distant from the other two branches of the AA family, namely Khasi-Khmuic and Mon-Khmer. The Khasi-Khmuic and Mon-Khmer are more similar to the Sino-Tibetan language. This is expected as our genetic data also confirms closer proximity and longer admixture of Khasi-Khmuic and Mon-Khmer speaking populations (Khasi and AAM) with Southern-EA populations. The admixture with populations of Northern EA-like ancestry is unique among the TB and their languages belong to Sino-Tibetan, a different language family altogether. In TBs this ancestry has been incorporated after the second migration wave. This leads us to conclude that the ancestral populations of TB have experienced a language shift, from a more proto-Khasi-Khmuic language to a language closer to that of the East Asians (the Tibeto Burman languages) and this has occurred due to the most recent admixture with populations with Northern EA-like ancestry.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://share.nibmg.ac.in/d/0b373e011a1d4f689cb2/, BMCB2021; https://www.nibmg.ac.in/fp/khasi_data.html, Khasi_data_2022.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Ethics Committee for Human Samples/Participants (IECHSP/2014/07), North-Eastern Hill University, Shillong, India. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AB and DT designed the study with the active participation of AC and PM. DT analyzed the data and prepared the final figures and tables. DT and AB wrote the manuscript with inputs from AC and PM. All authors read and approved the final manuscript.

Funding

This work was supported by the Department of Science and Technology (DST), Government of India. The infrastructure was provided by NIBMG and supported by the Department of Biotechnology (DBT), Government of India.

Acknowledgments

The authors are thankful to Diptarup Nandi at the National Institute of Biomedical Genomics for his valuable comments and suggestions. We also thank Genomics lab of NIBMG for the sequencing of the Khasi genomes and Arnab Ghosh for initial processing of the sequence data.

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.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1023870/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, 1655–1664. doi:10.1101/gr.094052.109

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., et al. (2003). Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res.* 13 (10), 2277–2290. doi:10.1101/gr.1413403

Basu, A., Sarkar-Roy, N., and Majumder, P. P. (2016). Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc. Natl. Acad. Sci. U. S. A.* 113 (6), 1594–1599. doi:10.1073/pnas.1513197113

Borkar, M., Ahmad, F., Khan, F., and Agrawal, S. (2011). Paleolithic spread of Y-chromosomal lineage of tribes in eastern and northeastern India. *Ann. Hum. Biol.* 38 (6), 736–746. doi:10.3109/03014460.2011.617389

Cavalli-Sforza, L. L., Piazza, L. L. C. S. P. M. A., Cavalli-Sforza, L., Menozzi, P., Piazza, A., and Princeton University Press (1994). *The history and geography of human genes*. Princeton, NJ, USA Princeton University Press.

Chazée, L. (1999). *The peoples of Laos: Rural and ethnic diversities: With an ethno-linguistic map*. Limited: White Lotus Company.Chennai, India (Thailand).

Cordaux, R., Saha, N., Bentley, G. R., Aunger, R., Sirajuddin, S. M., and Stoneking, M. (2003). Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur. J. Hum. Genet.* 11 (3), 253–264. doi:10.1038/sj.ejhg.5200949

Cordaux, R., Weiss, G., Saha, N., and Stoneking, M. (2004). the Northeast Indian passageway: A barrier or corridor for human migrations? *Mol. Biol. Evol.* 21 (8), 1525–1533. doi:10.1093/molbev/msh151

Das, B. M. (1970). Somatic variation among the Khasi populations of Assam, India. *Zmorph_anthropol.* 3, 259–266. doi:10.1127/zma/62/1970/259

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9 (2), 179–181. doi:10.1038/nmeth.1785

Diffloth, G. (2005a). "The contribution of linguistic palaeontology and Austroasiatic," in *The peopling of east Asia: Putting together archaeology, linguistics and GeneticsRoger blench and alicia sanchez-mazas laurent sagart* (Newyork, NY, USA: Routledge Curzon), 77–80.

Diffloth, G. (2005b). The peopling of east asia: putting together archaeology. *The contribution of linguistic paleontology to the homeland of Austro-asiaticlinguistics Genet.* 1, 79–82.

Endicott, P., Metspalu, M., and Kivisild, T. (2007). "Genetic evidence on modern human dispersals in South Asia: Y chromosome and mitochondrial DNA perspectives: The world through the eyes of two haploid genomes," in *The evolution and history of human populations in south Asia* (Springer), NewYork, NY, USA, 229–244.

Fuller, D. Q., Sato, Y.-I., Castillo, C., Qin, L., Weisskopf, A. R., Kingwell-Banham, E. J., et al. (2010). Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeol. Anthropol. Sci.* 2 (2), 115–131. doi:10.1007/s12520-010-0035-y

Gadgil, M., Shambu Prasad, U. V., Manoharan, S., and Patil, S. (1993). *Peopling of India*. Chennai, India: IHC.

Gayden, T., Mirabal, S., AliciaCadenas, M., Lacau, H., M Simms, T., Morlote, D., et al. (2009). Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. *J. Hum. Genet.* 54 (4), 216–223. doi:10.1038/jhg.2009.14

Guo, J., Wang, W., Zhao, K., Li, G., He, G., Zhao, J., et al. (2022). Genomic insights into Neolithic farming-related migrations in the junction of east and southeast Asia. *Am. J. Biol. Anthropol.* 177 (2), 328–342. doi:10.1002/ajpa.24434

GurdonThornhaghand Philip Richard (1914). *The Khasis*.NewYork, NY, USA Macmillan.

Higham, C. (2013). Hunter-gatherers in Southeast Asia: From prehistory to the present. *Hum. Biol.* 85 (1/3), 21–43. doi:10.3378/027.085.0302

Higham, C. F. (2017). First farmers in Mainland southeast Asia. *J. Indo-Pacific Archaeol.* 41, 13–21. doi:10.7152/jipa.v41i0.15014

Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., et al. (2006). Phylogeography and ethnogenesis of aboriginal southeast Asians. *Mol. Biol. Evol.* 23 (12), 2480–2491. doi:10.1093/molbev/msl124

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., et al. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* 72 (2), 313–332. doi:10.1086/346068

Krithika, S, Trivedi, R., Kashyap, V. K., Vasulu, T. S., and Kashyap, V. K. (2005). Genetic diversity at 15 microsatellite loci among the Adi Pasi population of Adi tribal cluster in Arunachal Pradesh, India. *Leg. Med.* 7 (5), 306–310. doi:10.1016/j.legalmed.2005.04.002

Krithika, S, Trivedi, R., Kashyap, V. K., Bharati, P., and Vasulu, T. S. (2006). Antiquity, geographic contiguity and genetic affinity among tibeto-burman populations of India: A microsatellite study. *Ann. Hum. Biol.* 33 (1), 26–42. doi:10.1080/03014460500424043

Krithika, S., Maji, S., and Vasulu, T. S. (2008). A microsatellite guided insight into the genetic status of Adi, an isolated hunting-gathering tribe of Northeast India. *PloS one* 3 (7), e2549. doi:10.1371/journal.pone.0002549

Langstieh, B. T., B Mohan Reddy, K. T., Kumar, V., and Singh, L. (2004). Genetic diversity and relationships among the tribes of Meghalaya compared to other Indian and Continental populations. *Hum. Biol.* 76, 569–590. doi:10.1353/hub.2004.0057

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* 361, 92–95. doi:10.1126/science.aat3188

Liu, C.-C., Witonsky, D., Gosling, A., Lee, J. H., Ringbauer, H., Hagan, R., et al. (2022). Ancient genomes from the Himalayas illuminate the genetic history of Tibetans and their Tibeto-Burman speaking neighbors. *Nat. Commun.* 13 (1), 1203–1214. doi:10.1038/s41467-022-28827-2

Longmailai, M. (2015). Language and culture in northeast india and beyond, 126.*Lexical and morphological resemblances of Khasi and dimasa*

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Douglas, C., Meehan, W., et al. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308 (5724), 1034–1036. doi:10.1126/science.1109792

Maity, B., Nunga, S. C., and Kashyap, V. K. (2003). Genetic polymorphism revealed by 13 tetrameric and 2 pentameric STR loci in four Mongoloid tribal population. *Forensic Sci. Int.* 132 (3), 216–222. doi:10.1016/s0379-0738(02)00436-x

Majumder, P. P. (2008). Genomic inferences on peopling of south Asia. *Curr. Opin. Genet. Dev.* 18 (3), 280–284. doi:10.1016/j.gde.2008.07.003

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93 (2), 278–288. doi:10.1016/j.ajhg.2013.06.020

Mastana, S. S., Murry, B., Sachdeva, M. P., Das, K., Young, D., Das, M. K., et al. (2007). Genetic variation of 13 STR loci in the four endogamous tribal populations of Eastern India. *Forensic Sci. Int.* 169 (2-3), 266–273. doi:10.1016/j.forsciint.2006.03.019

Matsumura, H. (2010). Quantitative and qualitative dental-morphology at man bac. Man bac. *Excav. a Neolithic Site North. Vietnam* 33, 43–63.

Matsumura, H. (2011). *Quantitative cranio-morphology at man bac. Man bac: The excavation of a late neolithic site in northern vietnam*, 21–32.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Peiros, I. (2011). Some thoughts on the problem of the Austro-Asiatic homeland. *J. Lang. Relatsh.* 6 (1), 101–114. doi:10.31826/9781463234119-009

Pickrell, J., and Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nat. Prec.* 1. 1. doi:10.1038/npre.2012.6956.1

Pinnow, H.-J., Kuiper, F. R. S., Greenberg, J. A. S., and Emeneau, M. (1942). The position of the Munda languages within the Austroasiatic language family. *Language* 18, 206.

Racimo, F., Sriram, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* 16 (6), 359–371. doi:10.1038/nrg3936

Reddy, B. M., Langstieh, B. T., Kumar, V., Nagaraja, T., Reddy, A. N. S., Meka, A., Reddy, A. G., et al. (2007). Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. *PLoS One* 2 (11), e1141. doi:10.1371/journal.pone.0001141

Salter-Townshend, M., and Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics* 212 (3), 869–889. doi:10.1534/genetics.119.302139

Shaun, P., 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 (3), 559–575. doi:10.1086/519795

Su, B., Xiao, C., Deka, R., Seielstad, M. T., Kangwanpong, D., Xiao, J., et al. (2000). Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum. Genet.* 107 (6), 582–590. doi:10.1007/s004390000406

Tagore, D., Aghakhanian, F., Naidu, R., Phipps, M. E., and Basu, A. (2021). Insights into the demographic history of Asia from common ancestry and admixture in the genomic landscape of present-day Austroasiatic speakers. *BMC Biol.* 19 (1), 61–19. doi:10.1186/s12915-021-00981-x

Thangaraj, K., Chaubey, G., Kivisild, T., Reddy, A. G., Singh, V. K., Rasalkar, A. A., et al. (2005). Reconstructing the origin of andaman islanders. *Science* 308 (5724), 996. doi:10.1126/science.1109987

Van Driem, G. (2005). *The peopling of east Asia: Putting together archaeology, linguistics and genetics*, Rouledge, England, UK, 81.Implications for population geneticists, archaeologists and prehistorians

Wang, L.-X., Lu, Y., Zhang, C., Wei, L.-H., Shi, Y., Huang, Y.-Z., et al. (2018). Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. *Mol. Genet. Genomics* 293 (5), 1293–1300. doi:10.1007/s00438-018-1461-2

Weir, B. S., and Cockerham, C. C. (1984). evolution, 1358–1370.Estimating F-statistics for the analysis of population structure

Yang, M. A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y.-C., Tsang, C.-h., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 369 (6501), 282–288. doi:10.1126/science.aba0909

Yu, X., and Hui, L. (2021). Origin of ethnic groups, linguistic families, and civilizations in China viewed from the Y chromosome. *Mol. Genet. Genomics.* 296 (4), 783–797. doi:10.1007/s00438-021-01794-x

TYPE Original Research
PUBLISHED 12 October 2022
DOI 10.3389/fgene.2022.946006

Population genetic characteristics of Hainan medaka with whole-genome resequencing

Zebin Yao[1], Shuisheng Long[1], Chun Wang[1], Chengqin Huang[1], Hairui Zhang[1], Liao Jian[1], Jingru Huang[1], Yusong Guo[1,2,3], Zhongdian Dong[1,2,3] and Zhongduo Wang[1,2,3]*

[1]Guangdong South China Sea Key Laboratory of Aquaculture for Aquatic Economic Animals, Fisheries College, Guangdong Ocean University, Zhanjiang, China, [2]Guangdong Provincial Engineering Laboratory for Mariculture Organism Breeding, Fisheries College, Guangdong Ocean University, Zhanjiang, China, [3]Guangdong Provincial Key Laboratory of Pathogenic Biology and Epidemiology for Aquatic Economic Animals, Fisheries College, Guangdong Ocean University, Zhanjiang, China

OPEN ACCESS

EDITED BY
Rosane Garcia Collevatti,
Universidade Federal de Goiás, Brazil

REVIEWED BY
Mingyou Li,
Shanghai Ocean University, China
Kiyoshi Naruse,
National Institute for Basic Biology, Japan
Shuying Li,
Zhejiang University, China

*CORRESPONDENCE
Zhongduo Wang,
wangzd@gdou.edu.cn

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

RECEIVED 17 May 2022
ACCEPTED 26 September 2022
PUBLISHED 12 October 2022

CITATION
Yao Z, Long S, Wang C, Huang C, Zhang H, Jian L, Huang J, Guo Y, Dong Z and Wang Z (2022), Population genetic characteristics of Hainan medaka with whole-genome resequencing.
Front. Genet. 13:946006.
doi: 10.3389/fgene.2022.946006

The *DMY* gene is deleted in all males of the Sanya population (SY-medaka) of the Hainan medaka, *Oryzias curvinotus*, as recently reported by us. However, due to limited knowledge regarding their population genetic background, it is difficult to explore the possible evolutionary pathway. Herein, we resequenced the whole genome of four populations, including SY-medaka. A total of 56 mitogenomes and 32,826,105 SNPs were identified. We found that the genetic differentiation is highest between SY-medaka and the other populations. The results of the population history of the *O. curvinotus* suggest that the SY-medaka has been in a bottleneck period recently. Further analysis shows that SY-medaka are the most strongly affected by environmental selection. Moreover, we screened some potential genomic regions, and the genes contained in these regions may explain the potential mechanism of the selection process of the SY-medaka. In conclusion, our study can provide new clues for the adaptation process of medaka in the new environment of Sanya.

KEYWORDS
***Oryzias curvinotus*, population genetics, genetics structure, population history, positive selection**

Introduction

The *Oryzias* has always been a species of interest for genetics and evolutionary biology, and research on the genetic diversity of *Oryzias* has been reported continuously. In terms of morphology and physiology, when the medaka was first studied, it was found that there were significant differences in the average number of anal fins of *Oryzias latipes* in different regions (Egami, 1954), while *O. latipes* in high latitudes grew faster than those in low latitudes (Yamahira and Takeshi, 2008). Genetic diversity among medaka has also

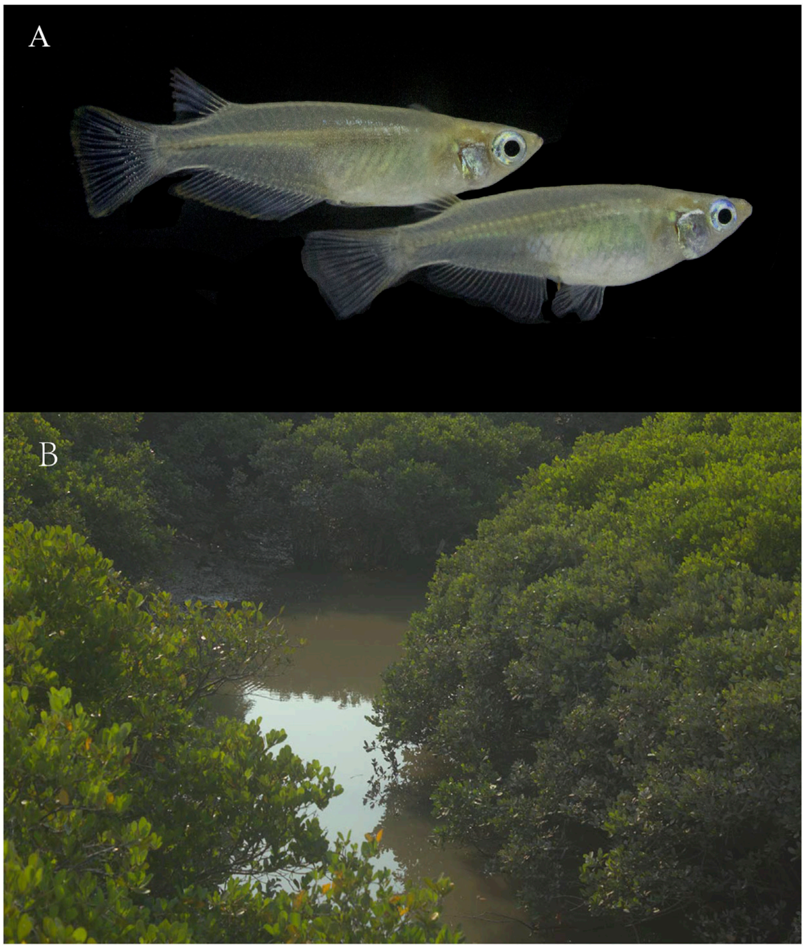


FIGURE 1
Oryzias curvinotus. **(A)** Male (upper) and female (lower) *Oryzias curvinotus* form Sanya. **(B)** Natural habitat of Hainan medaka.

been observed at the cytogenetic level; for example, there are differences in karyotypes among related species (Parenti, 2008). In addition, it was also found that *Oryzias* has two different sex determination systems: XX/XY and ZZ/ZW (Naruse et al., 2011). However, most relevant reports focus on *O. latipes*, the first reported teleost with a male sex-determining gene, *dmy* (Hirayama et al., 2010; Spivakov et al., 2014; Katsumura et al., 2019; Sutra et al., 2019).

As a closely related species of *O. latipes*, *O. curvinotus* inhabits mangrove forests along the coast of the South China Sea (Figures 1, 2). Moreover, *O. curvinotus* is the second species in genus *Oryzias* containing *dmy* (Matsuda et al., 2003). Especially, we have found that males of the Sanya population (SY-medaka) from Hainan Island generally lack *dmy* gene, which is obviously different from mainland populations such as the Gaoqiao population (GQ-medaka) (Dong et al., 2021). In addition, some phenomena of physiological differences have also been observed, for example, the population differences in the speeds of body color change (Supplementary Figure S2, unpublished) and the inconsistent growth rates (Wang, 2020) (Supplementary Figure S3). Obviously, the SY population of *O. curvinotus* may be different from other populations both in physiological and genetic levels. In particular, geographically, the SY-medaka is separated from the mainland population by the Qiongzhou Strait, which might hinder the gene flows of medaka because of the small body size, the low migrating ability and the bunched eggs by the filaments. Additionally, the climatic conditions also are different, such as water temperature by the latitudinal differences (Supplementary Figure S1). Thus, it is easy to associate external environmental factors with the effects on the various geographic population of the *O. curvinotus*. However, the lack of knowledge regarding population genetic background makes it difficult to reveal the biological mechanism of the above phenomena in *O. curvinotus*. Therefore, there is an urgent need to investigate the available gene pool and population genetics of *O. curvinotus*.

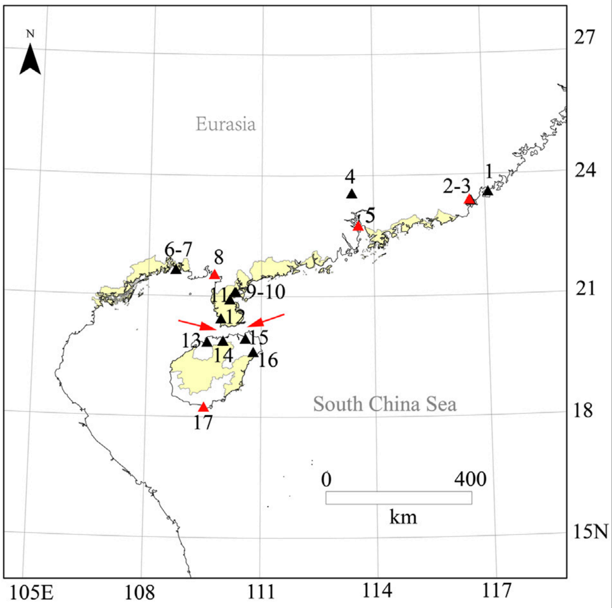


FIGURE 2
Species distribution of *O. curvinotus*. Species distribution map for *O. curvinotus* and some new investigation sites; 1, Raoping; 2, Dahao; 3, Niutianyang (ST); 4, Conghua; 5, Nansha (JWM); 6, Zhonggui; 7, Yamchow; 8, Gaoqiao (GQ); 9, Huguang; 10, Donghai dao; 11, Fucheng; 12, Leizhou; 13, Lingao County; 14, Chengmai; 15, Dongzhaigang; 16, Wenchang; 17, Sanya (SY); The location IDs are in common with Supplementary Table S2; Four collection sites of *O. curvinotus* for WGS are marked with red triangles. The red arrow refers to the Qiongzhou Strait.

Previous studies of genetic diversity and population structure have often been based on mitochondrial genome (mtDNA) (Hirayama et al., 2010; Mokodongan and Yamahira, 2015). But mtDNA was considered insufficient to fully describe population structure and history because of its single locus (Katsumura et al., 2019). Single nucleotide polymorphisms (SNPs) are considered an effective genetic marker for researching the genetic diversity and adaptive evolution of species. In recent years, as sequencing technology has become more mature and economical, the nuclear genomes of a large number of species have been successfully assembled. Whole genome sequencing (WGS) and SNPs detection have also been widely used. In vertebrates, SNPs occur almost once in every 1,000 nucleotides on average in human DNA, and some are strongly correlated with traits or genetic diseases (Kujovich, 2011; Ji et al., 2012; Ezzeddini et al., 2021). In addition, SNPs in many animals such as fish are associated with environmental adaptation. Recently, many potential genes related to environmental adaptation in fish have been discovered by using SNPs (Jones et al., 2012; Araki et al., 2018; Gaither, Gkafas, De Jong, et al., 2018; Narum et al., 2018; Onainor, 2019; Cádiz et al., 2020), which can usually provide some new clues for species evolution and environmental adaptation research.

Here, we performed a batch of whole-genome re-sequencing of the wild *O. curvinotus* from four geographic groups at different latitudes with the aim of obtaining a high density of SNPs and also analyzed the population structure and evolutionary history of *O. curvinotus* at the nuclear gene level.

Materials and methods

Sampling and sequencing

All the 56 wild individuals sequenced in this study were collected from four geographical locations (Figure 2, Supplementary Table S2) along the south coast of China with mangroves, of which Shantou (ST, E116.57° N23.38°), Jiuwangmiao (JWM, E113.58° N22.76°), and Gaoqiao (GQ, E109.74° N21.55°) are located in Guangdong Province, and Sanya (SY, E109.50° N18.25°) is located in Hainan Province.

After collection of each population, all the samples were euthanized and stored in absolute ethanol. Genomic DNA was extracted using the phenol-chloroform method. The fragment size of the sequencing library was 350 bp, and sequencing was conducted on the Illumina HiSeq Xten platform using terminal pairing (PE) 150 bp reads. The sequencing depth of each sample was ×10. The resequencing raw data of all samples in this study have been uploaded to the NCBI Sequence Reading Archive (SRA) with storage number SRR17331594-SRR17331649.

Read mapping and variant calling

Before mapping the resequenced data to the reference genome, we filtered the adapter sequence and low-quality reads of all original data to obtain high-quality clean data. BWA software (parameters: mem, -T 4, -K 32, -M) was used to compare sequencing results to the reference genome (GQ-medaka, Sequences have been submitted to GenBank, PRJNA821560) of *O. curvinotus*, and GATK4 (MarkDuplicates) software was used to remove PCR duplication (McKenna et al., 2010; Li, 2013).

The standard process of GATK4 software was used to identify variation and extract SNPs. High quality SNP data was filtered using GATK4 (VariantFiltration, default parameters) and VCFtools (parameters: -max-missing 0.5, -maf 0.01, -minDP 3, -minQ 30) (Danecek et al., 2011).

Annotation

Gene-based SNP annotation was performed according to the annotation of the *O. curvinotus* genome using the package

ANNOVAR (Wang et al., 2010). Based on the genome annotation, SNPs were categorized as occurring in exonic regions (overlapping with a coding exon), intronic regions (overlapping with an intron), splice sites (within 2 bp of a splicing junction), upstream and downstream regions (within a 2 kb region upstream or downstream from the transcription start site), or intergenic regions. SNPs in coding exons were further grouped as either synonymous SNPs or nonsynonymous SNPs. Additionally, mutations causing gain or loss of a stop codon were also classified as nonsynonymous SNPs. It should be noted that the statistical results for SNP classification may be redundant due to the overlap of genes.

Phylogenetic tree

The P-distance matrix was calculated based on the all SNPs data by VCF2Dis (https://github.com/BGI-shenzhen/VCF2Dis) and the NJ-tree was built with 1000 bootstrap by PHYLIP 3.69 (http://evolution.genetics.washington.edu/phylip.html).

Mitochondrial genome assembly and mtDNA tree

The open source toolkit, GetOrganelle (Jin et al., 2018) were used to assemble the mitochondrial genome based on the reads from the resequenced datasets. Each dataset of 56 samples was completely assembled into a circular sequence. The assembly results were confirmed to be accurate by re-mapping to the assembled mitochondrial sequence. Mitochondrial sequences of other species of the *Oryzias* were obtained from NCBI (Supplementary Table S3), multiple alignment were performed using the muscle method, and the maximum likelihood (ML) tree was built with 1000 bootstrap by MEGA (Kumar et al., 2018).

Principal component analysis

All SNPs of 56 individuals were used for principal component analysis by GCTA software (Yang et al., 2011). The first three important principal components were retained and graphed. The discrete points reflect the real structure of the population to a certain extent.

Population structure

To analyze the population structure, the program Admixture was used to estimate ancestry number in a model-based manner from all SNP genotype datasets. To explore the convergence of individuals, we predefined the number of genetic clusters from K = 3 to K = 6.

Linkage disequilibrium analysis

To estimate and compare the pattern of linkage disequilibrium (LD) of each population, the squared correlation coefficient (r^2) values between any two SNPs within 100 kb were computed by using the software PopLDdecay (Zhang et al., 2019). For plotting the LD attenuation curve, the SNP data of 56 individuals were divided into four groups according to geographical location, and the average r^2 of each group was calculated.

Demographic history reconstruction

PSMC software was used to construct the historical dynamics of the four populations of *O. curvinotus* (Li and Durbin, 2011). PSMC is based on the results of individual resequencing alignment and considers all heterozygous loci of an individual, not just SNPs. When estimating the population historical dynamics of *O. latipes*, it is assumed that the doubling time of each generation is 0.67 years and the mutation rate is 2.5×10^{-8} (Spivakov et al., 2014). However, the estimation results of our data based on these parameters are not very reliable. Therefore, in this study, to ensure that all possible results are considered, we used as many parameters as possible for estimation. According to the sexual maturity time of *O. curvinotus* and field investigation, multiple values were used for the mutation rate, ranging from 0.25×10^{-8} to 2.5×10^{-8}, and the generation intervals used were 0.25, 0.33, 0.5, 0.66, or 1 year. Default values were taken for other parameters.

Polymorphism levels and selection analyses

F_{ST} values between populations were calculated using vcftools software. In order to identify potential selected genes, we counted the θπ ratios and F_{ST} values in 40 kb windows in 20 kb steps. For comparing groups, the regions with maximum F_{ST} values (top 5%, as outliers) and maximum θπ ratio (top 5%, as outliers) were identified as selected regions. We used clusterProfiler to preform KEGG enrichment analysis on the candidate selective gene located in selected regions.

Results

Sequencing results and single nucleotide polymorphism identification and annotation

In this research, *O. curvinotus* from the four geographical groups are abbreviated as ST-medaka, JWM-medaka, GQ-

TABLE 1 Statistics for SNP annotation results.

Category	Number of SNPs
Exonic	875,581
Splice site	1,853
Upstream (2 kb)	1,346,688
Downstream (2 kb)	1,171,335
Upstream/downstream	325,702
Intergenic	13,595,255
Intronic	14,750,035
3′ UTR	574,339
5′ UTR	184,901
3′UTR/5′UTR	414
Category of exon	Number of SNPs
Synonymous	509,253
Nonsynonymous	355,894
Stop-gain	3,796
Stop-loss	683

medaka, and SY-medaka. Genome mapping of all individuals across the *O. curvinotus*'s 800 Mb genome resulted in an average of 98.21% sequencing coverage (Supplementary Table S1). Through identification and screening, we identified a total of 32,826,105 SNPs. Among them, 875,581 loci were located in exonic regions, 509,253 were synonymous variations, and 355,894 were nonsynonymous variations. Other basic statistical results are shown in Table 1.

Population genetic structure and phylogeny

The population genetic structure (Figure 3A) shows the population structure of the 56 individuals for population numbers (K) from 1 to 6. When K is equal to 2, 3 continental population are grouped, SY-medaka was separated. When K is equal to 3 or 4, ST-medaka and JWM-medaka are grouped. Although JWM-ST and JWM-GQ are almost equal in geographical straight-line distance, JWM-medaka was grouped with GQ-medaka. When K is equal to 4 or 5, JWM-medaka and ST-medaka are forced to be separated. When K is 4 or 6, the ancestral sequences of SY-medaka are separated, while the GQ-medaka can always be clearly separated. Statistical support for the different number of clusters was evaluated based on fivefold cross-validation implemented in Admixture (Supplementary Figure S4). The grouping of subgroups at the CV-error minimum is relatively reasonable. Based on the value at the CV-error minimum, the four populations of *O. curvinotus* are divided into three subgroups.

The VCF2Dis software was used to calculate the genetic distance between all individuals, and the neighbor-joining (NJ) method was used to construct the evolutionary tree. The unrooted tree (Figure 3B, Supplementary Figure S9) demonstrates that the genetic distance between ST-medaka and JWM-medaka is the shortest, and their genetic distance with SY-medaka is the longest, followed by their distance to GQ-medaka. Based on the complete mitochondrial sequence, we constructed the ML tree (Figure 4). Here, other species of *Oryzias* were used as the outgroup. Apart from the clustering of one individual (ST 16) that was different from the NJ tree based on SNPs, all other individuals formed clusters with other local individuals. This demonstrates that the genetic structure analysis based on SNPs is reliable. In addition, the result of ML tree indicate that *O. curvinotus* first differentiated into two branche. One branch includes ST-medaka and JWM-medaka, and the other branch includes GQ-medaka and SY-medaka.

The first three principal components, namely first principal component (PC1), second principal component (PC2), third principal component (PC3) were extracted and plotted. In particular, the PC1 (variance explained = 48.44%) separated the SY-medaka from the other populations, the PC2 (variance explained = 11.93%) separated the GQ-medaka from the other populations (Figure 3C). The PC3 (variance explained = 2.70%) indicated that ST-medaka and JWM-medaka were separated (Figure 3D). The PC3 were extracted and plotted. The variance interpretation sum of the first three principal components was 63.07%.

Population history of *O. curvinotus*

In Figure 5, we show the estimation results of historical population dynamics for different possible generation intervals and nucleic acid mutation rates. The results show that the historical dynamic results of most *O. curvinotus* are similar, and some poor results only appear for more extreme parameter values. Taking the estimate of $g = 0.67$, $\mu = 2.5 \times 10^{-8}$ as an example, the effective population sizes of all subgroups of medaka were basically the same until about 100,000 years ago, when the common ancestor of them had not yet diverged. Therefore, the effective population sizes dating back to this period are basically the same. During the period from about 40,000 to 100,000 years ago, the population size changes showed two trends, one is an upward trend containing SY-medaka and GQ-medaka, and the other is a downward trend containing JWM-medaka and ST-medaka, suggesting that two branches of the medaka diverged during this period. While the trends in effective population sizes of WM-medaka and ST-medaka remained consistent and the population sizes were very similar from approximately 10,000 to 40,000 years ago. In contrast, the trends of the other two populations were clearly distinguished, with GQ-medaka showing an upward and then a downward trend, while SY-medaka showed a sharp decline. All four populations of the *O. curvinotus* live along the coast of the South China Sea, but the demographic

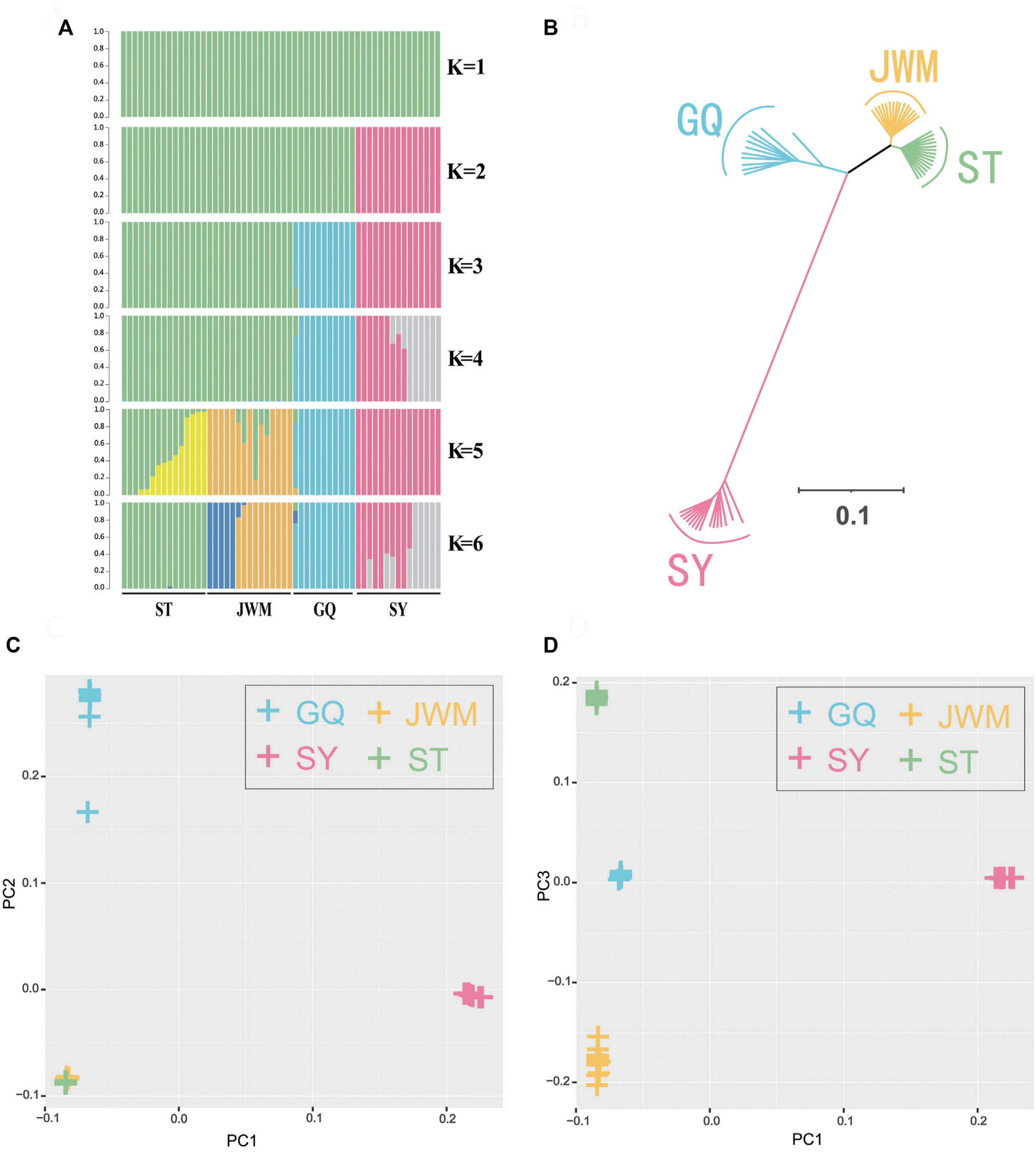


FIGURE 3
Population genetics. **(A)** Genetic structure of the *O. curvinotus* as inferred by Admixture analysis. The number of populations (K) from 1 to 6 is shown. Each color represents a different hypothetical ancestor. **(B)** Unrooted tree generated by the neighbor-joining method with 1000 bootstrap. **(C,D)** The principal component analysis (PCA) of all individuals.

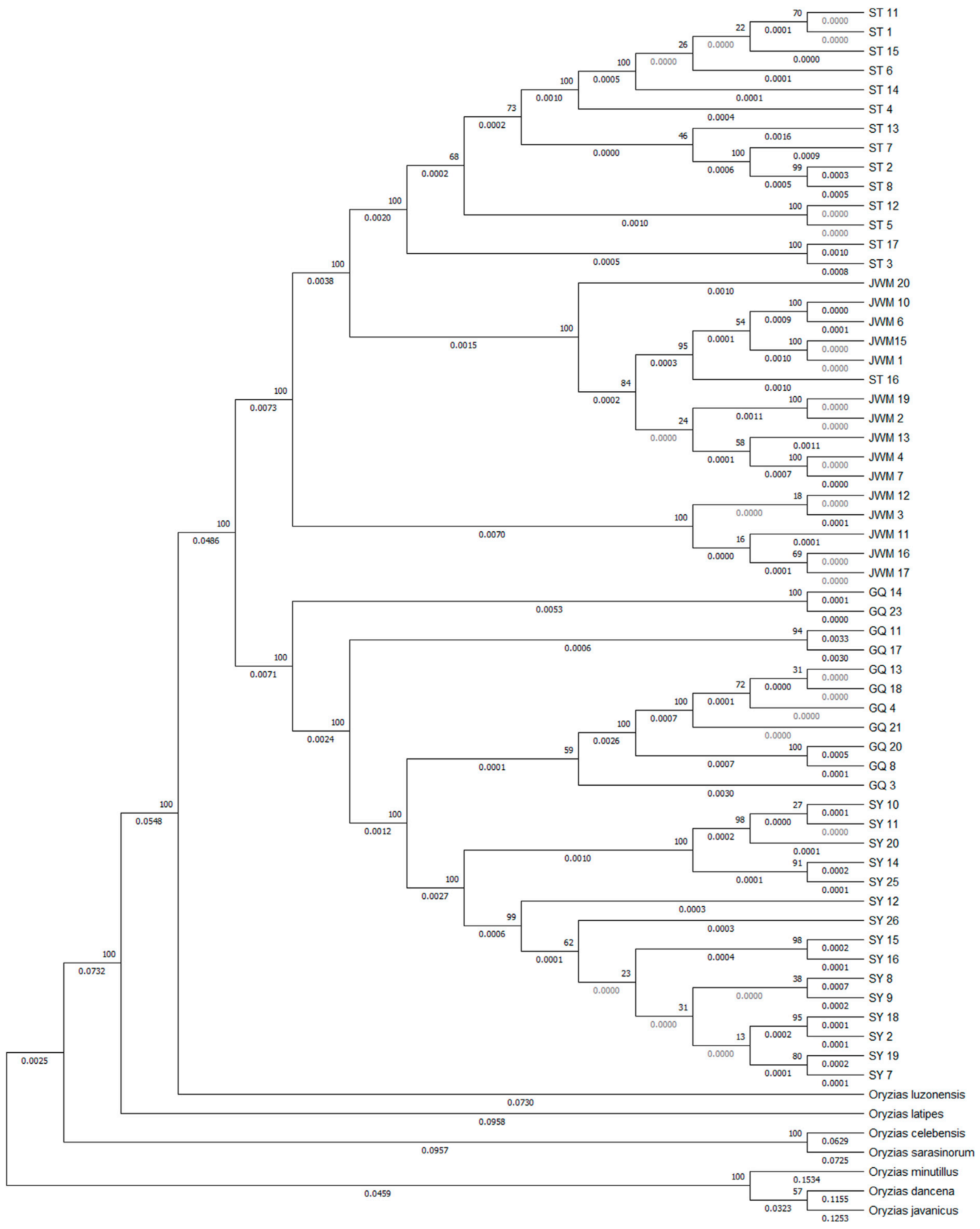


FIGURE 4
Based on the ML tree of mitochondrial whole sequences, some other fish of the *Oryizas* genus were used as outgroups. The bootstrap value (above branch) and branch lengths (below branch) are shown here.

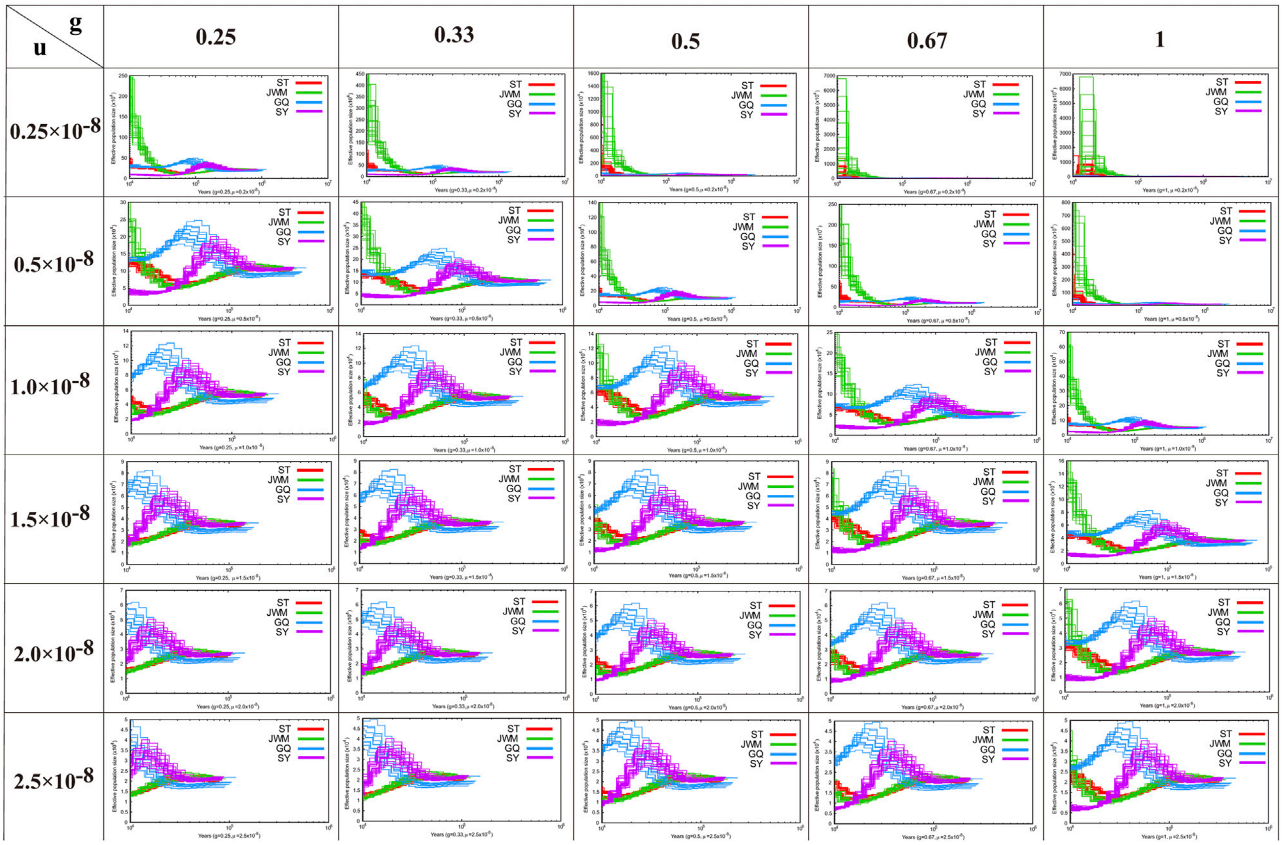


FIGURE 5
The results of the analysis of the demographic history of the *O. curvintous* are presented. Multiple values ranging from 0.25×10^{-8} to 2.5×10^{-8} were used for mutation rates (u). The generation intervals (g) were 0.25, 0.33, 0.5, 0.66, and 1 year based on the sexual maturity time of the medaka and field surveys.

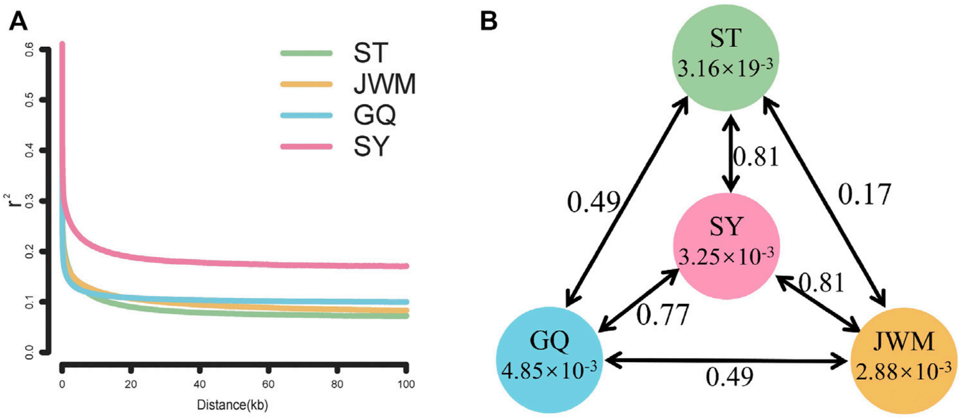


FIGURE 6
(A) Linkage disequilibrium patterns of four populations. **(B)** F_{ST} values among populations and θπ in each population.

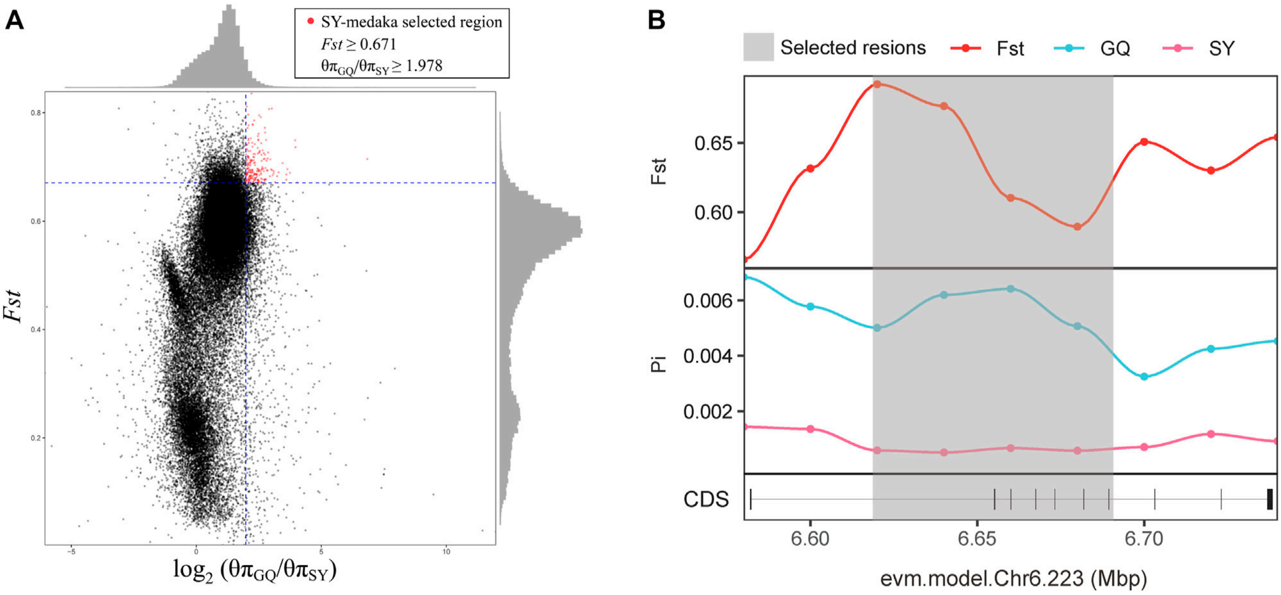


FIGURE 7
(A) The distribution of the θπ ratios (θπGQ-medaka/θπSY-medaka) and F_{ST} values (GQ-SY), calculated in 20-kb windows sliding in 10-kb step. Data points on the right of the vertical dashed line (corresponding to the 5% left tail of the empirical θπ ratio distribution), and above the horizontal dashed line (5% right tail of the empirical F_{ST} distribution) were identified as selected regions for SY-medaka (red points). **(B)** Examples of genes with strong selective sweep signals in GQ-medaka and SY-medaka. F_{ST} and θπ values are plotted using a 10-kb sliding window. Shaded genomic regions were the regions with strong selective signals for SY-medaka.

changes of several of them are very different, which may be due to the different effects of geological events and climatic environmental changes on them.

Linkage disequilibrium analysis

The LD attenuation diagram (Figure 6A) shows that the LD attenuation speed of each *O. curvinotus* population is very fast and stabilized at an attenuation distance of about 20 kb. In addition, the decay rates of the ST-medaka, JWM-medaka, and GQ-medaka are similar, but significantly lower than that of SY-medaka, indicating that SY-medaka may be under stronger selection pressure. It is worth noting that the LD value of the final convergence of the four groups is correlated with the latitude. At lower latitudes, the stable LD value of each group was higher.

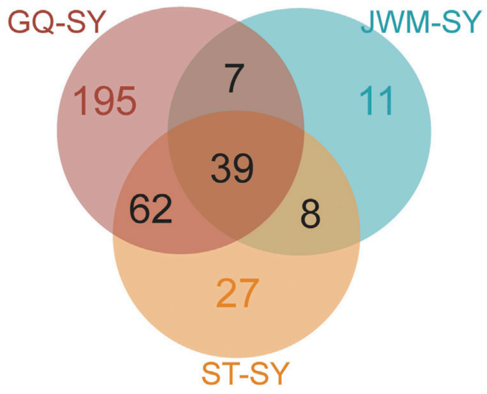


FIGURE 8
Venn diagram showing the intersection of the number of genes subject to selection in SY-medaka relative to other populations.

$\theta\pi$ and differentiation index (F_{ST})

Genome-wide scans were performed using the sliding-window approach, then θπ and F_{ST} were calculated between the populations (Supplementary Figures S5, S6). The results showed that SY-medaka had the highest F_{ST} value of 0.81 with the other populations, and the smallest F_{ST} between the ST-medaka and JWM-medaka (Figure 6B). In addition, GQ-medaka had the highest θπ of 4.85 × 10^{-3}, SY-medaka was the second highest at 3.25 × 10^{-3}, ST-medaka and JWM-medaka were the least at 3.16 × 19^{-3} and 2.88 × 10^{-3}, respectively. The results indicated that SY-medaka

TABLE 2 Functional pathway enrichment for selected genes of SY-medaka.

Gene name	Gene ID in genome	Gene annotation	KEGG path
ATF7IP	EVM.Model.Chr1.310	Activating transcription factor 7-interacting protein 1	—
txnl4a	EVM.Model.Chr11.1183	Thioredoxin-like protein 4A	—
KAF6729704.1	EVM.Model.Chr11.1185[a]	-	—
trit1	EVM.Model.Chr11.1188	tRNA dimethylallyltransferase	—
mycl	EVM.Model.Chr11.1189	Protein L-Myc-1b	—
mfsd2a	EVM.Model.Chr11.1190	Sodium-dependent lysophosphatidylcholine symporter 1-B	—
stk3	EVM.Model.Chr11.497	Serine/threonine-protein kinase 3	—
ridA	EVM.Model.Chr11.498	2-iminobutanoate/2-iminopropanoate deaminase	—
CYCS	EVM.Model.Chr11.499	—	—
rpl30	EVM.Model.Chr11.500	60S ribosomal protein L30	Ribosome, Coronavirus disease
LAPTM4B	EVM.Model.Chr11.501	Lysosomal-associated transmembrane protein 4B	Lysosome
RRM2B	EVM.Model.Chr11.502	Ribonucleoside-diphosphate reductase subunit M2	Purine metabolism Pyrimidine metabolism Glutathione metabolism Drug metabolism, p53 signaling pathway DNA Repair and Recombination Proteins
DTNBP1	EVM.Model.Chr11.508	Dysbindin	Membrane trafficking
RVE67048.1	EVM.Model.Chr11.514[a]	—	—
CEP192	EVM.Model.Chr16.896	Centrosomal protein of 192 kDa	—
Ankyrin-3-like isoform X1	EVM.Model.Chr19.33	Ankyrin-3	—
TNPO3	EVM.Model.Chr23.720	Transportin-3	Nucleocytoplasmic transport, Transfer RNA biogenesis
IRF5	EVM.Model.Chr23.721	Interferon regulatory factor 5	Toll-like receptor signaling pathway, Transcription factors
IRF5	EVM.Model.Chr23.722	Interferon regulatory factor 6	Toll-like receptor signaling pathway, Transcription factors
XP_011491718.1	EVM.Model.Chr23.723[a]	—	—
CDHR5-like isoform X1	EVM.Model.Chr23.724	Cadherin-related family member 5	Cell adhesion molecules
D (4) dopamine receptor-like	EVM.Model.Chr23.725	D (4) dopamine receptor	Neuroactive ligand-receptor interaction, Dopaminergic synapse, G-Protein Coupled Receptors
MBC8529904.1	EVM.Model.Chr23.726[a]	—	—
TSPAN12	EVM.Model.Chr23.745	Tetraspanin-12	—
ING3	EVM.Model.Chr23.746	Inhibitor of growth protein 3	Chromosome and associated proteins
CPED1	EVM.Model.Chr23.747	Cadherin-like and PC-esterase domain-containing protein 1	—
BUB1B-like isoform X1	EVM.Model.Chr24.309	Mitotic checkpoint serine/threonine-protein kinase BUB1 beta	—
SPINT1-like	EVM.Model.Chr24.310	Kunitz-type protease inhibitor 1	—
ARHGAP18	Evm.Model.Chr24.798	Rho GTPase-activating protein 18	—
XP_011490382.1	EVM.Model.Chr24.799	—	—
TMEM244	EVM.Model.Chr24.800	Transmembrane protein 244	—
	EVM.Model.Chr3.434	Liprin-beta-2	—

(Continued on following page)

TABLE 2 (*Continued*) Functional pathway enrichment for selected genes of SY-medaka.

Gene name	Gene ID in genome	Gene annotation	KEGG path
Liprin-beta-2-like isoform X1			
PPFIBP2	EVM.Model.Chr6.215	Liprin-beta-2	—
KAF6737022.1	EVM.Model.Chr6.216[a]	—	—
ARNTL protein 1	EVM.Model.Chr6.217	Aryl hydrocarbon receptor nuclear translocator-like protein 1	Dopaminergic synapse, Circadian rhythm, Transcription factors
GALNT18-like isoform X1	EVM.Model.Chr6.223	Polypeptide N-acetylgalactosaminyltransferase 18	—
SHANK3 isoform X1	EVM.Model.Chr6.915_EVM.Model.Chr6.916_EVM.Model.Chr6.917	SH3 and multiple ankyrin repeat domains protein 3	Glutamatergic synapse
ATF7IP2	EVM.Model.Chr8.986	Activating transcription factor 7-interacting protein 1	—
EMP2	EVM.Model.Chr8.987	Epithelial membrane protein 2	—

Note: —, indicates no relevant information.

[a]Indicates that it cannot be annotated, we provide the access number with the highest score by blastx in the corresponding "gene name" column.

produced a large genetic differentiation. Furthermore, that of SY-medaka depicted a highly genetically differentiated population from other populations, suggesting that SY-medaka may have been subjected to strong selection.

Selected gene

In order to identify potential selected genes in SY-medaka, and to detect regions with significant signatures of a selective sweep, we considered the distribution of the θπ ratios and F_{ST} values. We selected windows simultaneously with significant high θπ ratios and significant high F_{ST} values of the empirical distribution as regions with strong selective sweep signals along the genome (Figure 7, Supplementary Figures S7). Through the intersection of the top 5% windows of F_{ST} values and θπ ratios, we screened some strong selection signals of the three subpopulations, in which 136 genes were obtained by comparing ST to SY (ST-SY), 65 genes were obtained for JWM-SY, 303 genes were obtained for GQ-SY, and 39 candidate genes were obtained by taking the intersection of the three groups of genes (Figure 8). 39 selected genes of SY-medaka participate in multiple biological processes, which may indicate the potential processes through which *O. curvinotus* have adapted to the new environment of Sanya. We found that the annotation information of six genes was incomplete, and their specific functions need to be further studied. We performed KEGG enrichment analysis on the other 33 candidate genes to further explore the potential mechanism by which *O. curvinotus* adapted to the new environment of Sanya (Table 2). KEGG enrichment results showed that 12 genes were enriched in KEGG pathways, and only three genes (EVM.Model.Chr6.217, EVM.Model.Chr23.722, and EVM. Model.Chr23.721) were significantly enriched in two pathways (ko04711: circadian rhythm—fly; ko04620: Toll-like receptor signaling pathway).

Discussion

Population genetic structure and genetic differentiation of *O. curvinotus*

We have assessed for the first time the population genetic structure and population history of the *O. curvinotus*, and found that the *O. curvinotus* can be divided into three subgroups, with the ST-medaka and JWM-medaka forming a subgroup located in the east, and the GQ-medaka and SY-medaka each forming a separate subgroup, which we suggest is the result of local subgroups adapting to their respective habitats.

We found that the genetic divergence of SY-medaka from other geographic groups is large, which may result from different environmental differences, and that such environmental differences are likely to be latitude-related. It has been shown that *O. latipes* in high latitudes grew faster than those in low latitudes (Yamahira and Takeshi, 2008). Similar results were observed in *O. curvinotus*, and other physiological indicators such as heart rate were also found to differ between geographic groups at different latitudes (Wang, 2020) (Supplementary Figures S8, unpublished). We also found that the genetic differentiation of F_{ST} between geographic groups showed some correlation with their latitudinal span (Figure 9), suggesting a genetic basis for the physiological differences between geographic groups at different latitudes. Interestingly, the rate of LD decay also shows some correlation with latitude. In the *O. curvinotus*, LD decays particularly rapidly, reaching a steady state at r^2 at about 20 kb, which is probably due to the very

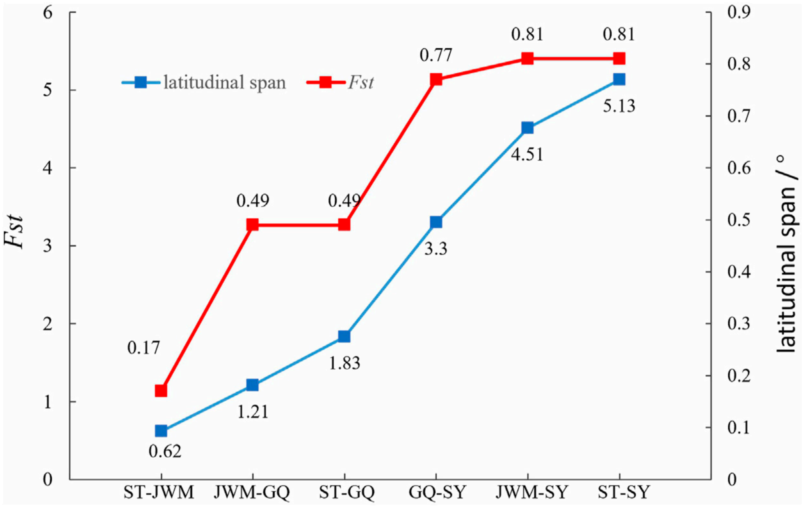


FIGURE 9
The relation between F_{ST} between geographic groups and their latitudinal span.

short generation interval (about 3–6 months). The same situation exists in *O. latipes* (Spivakov et al., 2014; Katsumura et al., 2019). This indicates that the fishes of *Oryzias* have some commonalities, which do not change with environmental factors.

In addition, SY-medaka is a special and interesting population. SY-medaka decayed more rapidly amongst the four geographic groups, significantly more than in the other groups, which we attribute to the effects of long-term environmental selection pressures on SY-medaka.

Population history and selective pressures of *O. curvinotus*

Previous studies have suggested that the latipes group (comprising *O. latipes*, *O. curvinotus* and *O. luzonensis*) arose in 27.3 Mya and that *O. latipes* formed in 9.2 Mya, on the basis of which the species formation and population divergence of the *O. curvinotus* would have occurred within 9.2 million years (Rabosky et al., 2018). Geological evidence suggests that Hainan Island, where SY-medaka is located, was initially closely linked to Eurasia. However, due to the influence of the Indian plate on Eurasia, Hainan Island separated and drifted to its present location. This process occurred at about 65 Mya-24 Mya (Guanghe, 2018), suggesting that Hainan Island was completely detached from the mainland by the time the *O. curvinotus* formed. The separation of Hainan Island indirectly led to the creation of the Qiongzhou Strait, a necessary precondition for the geographic isolation of SY-medaka from other groups, however this geographic isolation was unstable and subject to sea-level elevation.

Some paleoceanographic data demonstrate that the sea level has been changing continuously throughout history, with the last drastic change occurring about 20,000 years ago (Ganopolski et al., 2010; Hertzberg, 2015; Peter et al., 2009). The global sea level continued to decline due to the advent of the Last Glacial Period, the coastline of the Eurasian continent moved southward, and the bottom of the Qiongzhou Strait emerged. The decline of the sea level not only completely separated the groups of *O. curvinotus*, but also turned the original water surface into land. Only the low-lying and river channels are still covered with water, which obviously restricts the living space of *O. curvinotus*. This is consistent with the historical changes that occurred in the populations we analyzed. Since about 100,000 years ago, each group has undergone population decline. During the Last Glacial Period, there was a decline in the population number of the four populations, but the population size of the JWM-medaka and GQ-medaka in Shantou recovered or even exceeded their prior numbers. However, the populations of SY-medaka and GQ-medaka, despite historically experiencing a short population increase, have been decreasing ever since, and both populations seemed to be in a bottleneck period. However, SY-medaka declined the most, to a much greater extent than GQ-medaka, and there was no trend of population recovery in many estimation results. Not only that, but we also found that the population size change of the Southern population of *O. latipes* during the Last Glacial Period was perfectly in line with our speculation that the population of Japanese Southern medaka started to decline at the onset of the Last Glacial Period and began to expand almost simultaneously with the end of the ice age (Spivakov et al., 2014), thus we suggest that the ice age affected the sea level and thus the population size of all *Oryzias* species.

Unfortunately, however, for our data, PSMC can only estimate up to the last 10,000 years, and it remains to be further explored whether GQ-medaka and SY-medaka will show expansion after the bottleneck, like the population groups or *O. latipes*.

At the same time, we found that the trends in population change of ST-medaka and JWM-medaka were almost the same, while the change trends of GQ-medaka and SY-medaka were almost the same, although their population levels differed greatly. The ML tree constructed based on mitochondria also reflects that GQ-medaka and SY-medaka are closely related, indicating that they have a common ancestor. However, the NJ tree based on nuclear SNPs revealed that GQ-medaka and SY-medaka are far from each other. Therefore, we believe that SY-medaka was subjected to strong selection that resulted in GQ-medaka and SY-medaka having a common ancestor while being genetically distant.

Potential environmental factors acting as selection pressures on SY-medaka

In terms of possible environmental factors acting as selective pressures on SY-medaka, we only considered temperature and ultraviolet light. Because *O. curvinotus* is mainly distributed along the coast of southern China, the climatic conditions of the four sampling points are almost the same, except that the light and temperature may have been different. Using climate data collected from the Weather Spark website from 1 January 1980, to 31 December 2016, we compared the climate and weather changes of Shantou, Guangzhou (representing Jiuwangmiao), Gaoqiao, and Sanya within 1 year ("Weather Spark, 2018.). While there were no regional differences in the average daily short-wave solar energy, differences were observed for water temperature. Except for the July–September period, the monthly average water temperature in Sanya was higher than in the other three geographical locations (Supplementary Figures S1), especially from January to February in the cold season, suggesting that the water temperature difference caused by different latitudes may be an important factor by which environmental selection for *O. curvinotus* occurs.

Selected genes of SY-medaka

39 selected genes of SY-medaka participate in multiple biological processes, which may indicate the potential processes through which *O. curvinotus* have adapted to the new environment of Sanya. Among them, the circadian rhythm pathway may provide possible clues for environmental stress factors. In mammals, heat stress can regulate circadian rhythm through a series of physiological reactions (Hertzberg, 2015), to make the body adapt to this thermal environment. In addition, UV stress research has found that clock genes can regulate the circadian clock to mediate sequential and hierarchical interactions between the heat shock response and tumor inhibition mechanisms, thereby protecting cells from UV stress (Kawamura et al., 2018). Of course, further experiments are needed to verify whether the *O. curvinotus* also responds to environmental changes through these mechanisms. Other genes may also be related to some potential mechanisms. For example, we found that many genes are closely related to cell growth and death. The expression of *stk3*, *ing3*, *trit1,* and *laptm4b* may inhibit cell growth and induce apoptosis (Golovko et al., 2000; Nagashima et al., 2003; Yarham et al., 2014; Blom et al., 2015; Wang et al., 2020), and *cep192* and *bub1b* may play an important role in mitosis (Chan et al., 1999; Zhu et al., 2008; Joukov et al., 2014). Further research is required to better understand their roles and possible mechanisms.

Although many genes lack annotation information and related pathway research, we believe that positive selection of these genes is very important for the adaptation of SY-medaka to the new environment. Furthermore, these genes provide us with an important research basis and research direction. We also note that the genome mapping rate of SY-medaka is about 1% lower than that of other populations, suggesting that more genes are selected against and not found in SY-medaka. It is difficult to identify these genes through the GQ-medaka genome. Therefore, we have started an SY-medaka genome project at the time of writing this paper, which is expected to further explore the environmental adaptation process of SY-medaka.

Conclusion

In this study, we performed WGS on 56 *O. curvinotus* individuals from four geographic locations, identified 32,826,105 SNPs. The population genetic structure analysis showed that the *O. curvinotus* can be divided into three subgroups. Among them, ST-medaka and JWM-medaka were the most closely related, while the SY-medaka was the most divergent from the other populations. The results of the population history of the *O. curvinotus* suggest that the SY-medaka has been in a bottleneck period recently and a variety of evidence shows that them were subjected to continuous strong selection pressure. By selective screening, we also identified some potential gene regions that were subject to significant positive selection in the SY-medaka. Candidate genes located in these regions may play important roles in biological processes such as circadian rhythm and cell cycle regulation, which may suggest potential adaptive processes that occurred in the new Sanya environment. Our study provides a genetic basis for the study of environmental adaptation of medaka.

Data availability statement

The resequencing raw data of all samples in this study have been uploaded to the NCBI Sequence Reading Archive (SRA) with storage number of PRJNA792300. Genome sequences have been submitted to GenBank (PRJNA821560). SNPs data for all samples have been submitted to Dryad, https://doi.org/10.5061/dryad.r2280gbf7. The complete mitochondrial sequence was submitted in fasta format in the Supplementary Material.

Ethics statement

The animal study was reviewed and approved by the Approval of Animal Use Protocol, Institutional Animal Care and Use Committees, Fisheries college of Guangdong Ocean University.

Author contributions

ZY: Conceptualization, Data curation, Supervision, Funding acquisition, Software, Methodology, Writing—original draft and Writing—editing. SL and CW: Data curation, Methodology, Resources, Software, and Visualization. CH, HZ, and LJ: Data curation, Formal analysis, Investigation, Resources, Software and Supervision. JH: Data curation, Formal analysis, Investigation. YG and ZD: Conceptualization, Formal analysis; Funding acquisition; Project administration, Resources, Supervision, Validation and Writing—original draft. ZW: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Software, Supervision and Writing—original draft.

Funding

This project was supported by Guangdong-Guangxi Joint Fund (No. 2020A1515410009) from Ministry of Science and Technology of China, the Natural Science Foundation of China (No. 31972794), National Natural Science Foundation of China Youth Fund (No. 41806195), the National Natural Science Foundation of Guangdong (No. 2022A1515011441).

Acknowledgments

We are grateful to Jin Gao and Benben Miao for their helpful input regarding data analysis.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.946006/full#supplementary-material

References

Araki, K., Aokic, J., Kawase, J., Hamada, K., Ozaki, A., Fujimoto, H., et al. (2018). Whole genome sequencing of greater amberjack (*Seriola dumerili*) for SNP identification on aligned scaffolds and genome structural variation analysis using parallel resequencing. *Int. J. Genomics* 2018, 7984292–7984312. doi:10.1155/2018/7984292

Blom, T., Li, S., Dichlberger, A., Bäck, N., Kim, Y. A., Loizides-Mangold, U., et al. (2015). LAPTM4B facilitates late endosomal ceramide export to control cell death pathways. *Nat. Chem. Biol.* 11 (10), 799–806. doi:10.1038/nchembio.1889

Cádiz, M. I., López, M. E., Díaz-Domínguez, D., Cáceres, G., Yoshida, G. M., Gomez-Uchida, D., et al. (2020). Whole genome re-sequencing reveals recent signatures of selection in three strains of farmed Nile tilapia (*Oreochromis niloticus*). *Sci. Rep.* 10 (1), 11514–14. doi:10.1038/s41598-020-68064-5

Chan, G. K. T., Jablonski, S. A., Sudakin, V., Hittle, J. C., and Yen, T. J. (1999). Human BUBR1 is a mitotic checkpoint kinase that monitors CENP-E functions at kinetochores and binds the cyclosome/APC. *J. Cell Biol.* 146 (5), 941–954. doi:10.1083/jcb.146.5.941

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, Z., Li, X., Yao, Z., Wang, C., Guo, Y., Wang, Q., et al. (2021). *Oryzias curvinotus* in Sanya does not contain the male sex-determining gene dmy. *Animals.* 11 (5), 1327. doi:10.3390/ani11051327

Egami, N. (1954). Geographical variations in the male characters of the fish, *Oryzias latipes*. *Annot. Zool. Jpn.* 27 (1), 7–12.

Ezzeddini, R., Somi, M. H., Taghikhani, M., Moaddab, S.-Y., Masnadi Shirazi, K., Shirmohammadi, M., et al. (2021). Association of Foxp3 rs3761548 polymorphism with cytokines concentration in gastric adenocarcinoma patients. *Cytokine* 138, 155351. doi:10.1016/j.cyto.2020.155351

Gaither, M. R., Gkafas, G. A., de Jong, M., Sarigol, F., Neat, F., Regnier, T., et al. (2018). Genomics of habitat choice and adaptive evolution in a deep-sea fish. *Nat. Ecol. Evol.* 2 (4), 680–687. doi:10.1038/s41559-018-0482-x

Ganopolski, A., Calov, R., and Claussen, M. (2010). Simulation of the last glacial cycle with a coupled climate ice-sheet model of intermediate complexity. *Clim. Past.* 6 (2), 229–244. doi:10.5194/cp-6-229-2010

Golovko, A., Hjälm, G., Sitbon, F., and Nicander, B. (2000)., 258. Netherlands, 85–93. doi:10.1016/s0378-1119(00)00421-2Cloning of a human tRNA isopentenyl transferase*Gene*1–2

Guanghe, L. (2018). A study of the Genesis of hainan island. *Geol. China* 45 (4), 693–705.

Hertzberg, M. (2015). Climate change reconsidered II — physical science. *Energy & Environ.* 26, 547–553. The Heartland Institute. doi:10.1260/0958-305x.26.3.547

Hirayama, M., Mukai, T., Miya, M., Murata, Y., Sekiya, Y., Yamashita, T., et al. (2010). Intraspecific variation in the mitochondrial genome among local populations of Medaka *Oryzias latipes*. *Gene* 457 (1–2), 13–24. doi:10.1016/j.gene.2010.02.012

Ji, G., Long, Y., Zhou, Y., Huang, C., Gu, A., and Wang, X. (2012). Common variants in mismatch repair genes associated with increased risk of sperm DNA damage and male infertility. *BMC Med.* 10 (1), 49. doi:10.1186/1741-7015-10-49

Jin, J. J., Yu, W. Bin, Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., et al. (2018). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *BioRxiv* 21 (1), 1–31.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484 (7392), 55–61. doi:10.1038/nature10944

Joukov, V., Walter, J. C., and De Nicolo, A. (2014). The Cep192-organized aurora A-Plk1 cascade is essential for centrosome cycle and bipolar spindle assembly. *Mol. Cell* 55 (4), 578–591. doi:10.1016/j.molcel.2014.06.016

Katsumura, T., Oda, S., Mitani, H., and Oota, H. (2019). Medaka population genome structure and demographic history described via genotyping-by-sequencing. *G3 Genes* 9 (1), 217–228. doi:10.1534/g3.118.200779

Kawamura, G., Hattori, M., Takamatsu, K., Tsukada, T., Ninomiya, Y., Benjamin, I., et al. (2018). Cooperative interaction among BMAL1, HSF1, and p53 protects mammalian cells from UV stress", *Communications Biology*. *Commun. Biol.* 1 (1), 204–209. doi:10.1038/s42003-018-0209-1

Kujovich, J. L. (2011). Factor V leiden thrombophilia. *Genet. Med.* 13 (1), 1–16. Official Journal of the American College of Medical Genetics, United States. doi:10.1097/GIM.0b013e3181faa0f2

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). Mega X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. Cambridge: ArXiv Preprint ArXiv:1303.3997.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475 (7357), 493–496. doi:10.1038/nature10231

Matsuda, M., Sato, T., Toyazaki, Y., Nagahama, Y., Hamaguchi, S., and Sakaizumi, M. (2003). *Oryzias curvinotus* has *DMY*, a gene that is required for male development in the medaka, *O. latipes*. *Zool. Sci.* 20 (2), 159–161. doi:10.2108/zsj.20.159

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110

Mokodongan, D. F., and Yamahira, K. (2015). Origin and intra-island diversification of Sulawesi endemic Adrianichthyidae. *Mol. Phylogenet. Evol.* 93, 150–160. doi:10.1016/j.ympev.2015.07.024

Nagashima, M., Shiseki, M., Pedeux, R. M., Okamura, S., Kitahama-Shiseki, M., Miura, K., et al. (2003). A novel PHD-finger motif protein, p47ING3, modulates p53-mediated transcription, cell cycle control, and apoptosis. *Oncogene* 22 (3), 343–350. doi:10.1038/sj.onc.1206115

Narum, S. R., Di Genova, A., Micheletti, S. J., and Maass, A. (2018). Genomic variation underlying complex life-history traits revealed by genome sequencing in Chinook salmon. *Proc. Biol. Sci.* 285 (1883), 20180935. doi:10.1098/rspb.2018.0935

Naruse, K., Tanaka, M., and Takeda, H. (2011). in *Medaka: A model for organogenesis, human disease, and evolution*. Editors N. Kiyoshi, T. Minoru, and T. Hiroyuki. 1st ed. (Tokyo: Springer Science & Business Media). available at. doi:10.1007/978-4-431-92691-7

Onainor, E. R. (2019). Neutral and adaptive drivers of genomic change in introduced brook trout (*Salvelinus fontinalis*) populations revealed by pooled whole-genome re-sequencing 1, 105–112.

Parenti, L. R. (2008). A phylogenetic analysis and taxonomic revision of ricefishes, Oryzias and relatives (Beloniformes, Adrianichthyidae). *Zoological J. Linn. Soc.* 154 (3), 494–610. doi:10.1111/j.1096-3642.2008.00417.x

Parenti, L. (2012). *The IUCN red list of threatened species 2012",* IUCN. en: RLTS.T181194A1708251. available at. doi:10.2305/IUCN.UK.2012-1

Peter, U., Clark, A. S., Dykejeremy, D., Shakunanders, E. C., Jorie, C., Barbara, W., et al. (2009). The last glacial maximum", *science*. *Am. Assoc. Adv. Sci.* 325 (5941), 710–714.

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., et al. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* 559 (7714), 392–395. doi:10.1038/s41586-018-0273-1

Setiamarga, D. H. E., Miya, M., Yamanoue, Y., Azuma, Y., Inoue, J. G., Ishiguro, N. B., et al. (2009). Divergence time of the two regional medaka populations in Japan as a new time scale for comparative genomics of vertebrates. *Biol. Lett.* 5 (6), 812–816. doi:10.1098/rsbl.2009.0419

Spark, Weather (2018). The weather year round anywhere on earth. https://weatherspark.com (Accessed December 1, 2021).

Spivakov, M., Auer, T. O., Peravali, R., Dunham, I., Dolle, D., Fujiyama, A., et al. (2014). Genomic and phenotypic characterization of a wild medaka population: Towards the establishment of an isogenic population genetic resource in fish. *G3 (Bethesda)* 4 (3), 433–445. doi:10.1534/g3.113.008722

Sutra, N., Kusumi, J., Montenegro, J., Kobayashi, H., Fujimoto, S., Masengi, K. W. A., et al. (2019). Evidence for sympatric speciation in a Wallacean ancient lake", *Evolution*. *Evolution* 73 (9), 1898–1915. doi:10.1111/evo.13821

Wang, C. (2020). *Comparison of phenotypes of three geographical populations of oryzias curvinotus and preliminary investigation of its sex determining genes"*. Zhangjiang: Guangdong Ocean University. master's thesis.

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16), e164. doi:10.1093/nar/gkq603

Wang, X., Wang, F., Zhang, Z. G., Yang, X. M., and Zhang, R. (2020). STK3 suppresses ovarian cancer progression by activating NF- κ B signaling to recruit cd8+t-cells. *J. Immunol. Res.* 2020, 7263602. available at. doi:10.1155/2020/7263602

Yamahira, K., and Takeshi, K. (2008). Variation in juvenile growth rates among and within latitudinal populations of the medaka. *Popul. Ecol.* 50 (1), 3–8. doi:10.1007/s10144-007-0055-3

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi:10.1016/j.ajhg.2010.11.011

Yarham, J. W., Lamichhane, T. N., Pyle, A., Mattijssen, S., Baruffini, E., Bruni, F., et al. (2014). Defective i6A37 modification of mitochondrial and cytosolic tRNAs results from pathogenic mutations in TRIT1 and its substrate tRNA. *PLoS Genet.* 10 (6), e1004424. doi:10.1371/journal.pgen.1004424

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., and Yang, T. L. (2019). PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35 (10), 1786–1788. doi:10.1093/bioinformatics/bty875

Zhu, F., Lawo, S., Bird, A., Pinchev, D., Ralph, A., Richter, C., et al. (2008). The mammalian SPD-2 ortholog Cep192 regulates centrosome biogenesis. *Curr. Biol.* 18 (2), 136–141. doi:10.1016/j.cub.2007.12.055

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fgene.2022.977517

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Shaoqing Wen,
Fudan University, China
Marcin Woźniak,
Nicolaus Copernicus University in
Toruń, Poland
Dengcheng Yang,
Beijing Forestry University, China

*CORRESPONDENCE
Horolma Pamjav,
phorolma@hotmail.com
Erzsébet Fóthi,
Fothi.Erzsebet@abtk.hu

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

RECEIVED 24 June 2022
ACCEPTED 03 October 2022
PUBLISHED 17 October 2022

CITATION
Pamjav H, Fóthi Á, Dudás D, Tapasztó A, Krizsik V and Fóthi E (2022), The paternal genetic legacy of Hungarian-speaking Rétköz (Hungary) and Váh valley (Slovakia) populations.
Front. Genet. 13:977517.
doi: 10.3389/fgene.2022.977517

The paternal genetic legacy of Hungarian-speaking Rétköz (Hungary) and Váh valley (Slovakia) populations

Horolma Pamjav[1]*, Ábel Fóthi[2], Dániel Dudás[1,3], Attila Tapasztó[1], Virág Krizsik[2] and Erzsébet Fóthi[2]*

[1]Department of Reference sample analysis, Institute of Forensic Genetics, Hungarian Institutes for Forensic Sciences, Budapest, Hungary, [2]Institute of Archaeogenomics, Research Centre for the Humanities, Budapest, Hungary, [3]Departmant of Genetics, Eötvös Lorand University, Budapest, Hungary

One hundred and six Rétköz and 48 Váh valley samples were collected from the contact zones of Hungarian-Slovakian territories and were genotyped for Y-chromosomal haplotypes and haplogroups. The results were compared with contemporary and archaic data from published sources. The genetic composition of the Rétköz population from Hungary and the Váh valley population from Slovakia indicates different histories. In the Rétköz population, the paternal lineages that were also found in the Hungarian Conquerors, such as R1a-Z93, N-M46, Q-M242, and R1b-L23, were better preserved. These haplogroups occurred in 10% of the population. The population of the Váh valley, however, is characterized by the complete absence of these haplogroups. Our study did not detect a genetic link between the Váh valley population and the Hungarian Conquerors; the genetic composition of the Váh valley population is similar to that of the surrounding Indo-European populations. The Hungarian Rétköz males shared common haplotypes with ancient Xiongnu, ancient Avar, Caucasian Avar, Abkhazian, Balkarian, and Circassian males within haplogroups R1a-Z93, N1c-M46, and R1b-L23, indicating a common genetic footprint. Another difference between the two studied Hungarian populations can be concluded from the Fst-based MDS plot. The Váh valley, in the western part of the Hungarian-Slovakian contact zone, is genetically closer to the Western Europeans. In contrast, Rétköz is in the eastern part of that zone and therefore closer to the Eastern Europeans.

KEYWORDS

Y haplotypes and haplogroups, Hungarian-speakers, Rétköz and Váh valley populations, human demographic history, Y-STRs and Y-SNPs

1 Introduction

The Carpathian Basin was historically the destination for several nomadic tribes that migrated westwards from Inner and Central Asia towards Europe. The ancient Hungarians (Steppe Magyars) entered the Carpathian Basin from the east in the ninth century CE and settled there (Róna-Tas, 1999). The genetic legacy of the populations in the Carpathian Basin can be studied only if the genetic structure of them is compared to that of ancient populations. Fortunately, in recent years, the number of aDNA genetic results based on both uniparental and whole genome data has increased in Hungary (Neparáczki et al., 2017a; Neparáczki et al., 2017b; Neparáczki et al., 2018; Olasz et al., 2018; Neparáczki et al., 2019; Csáky et al., 2020; Fóthi et al., 2020; Maár et al., 2021; Nagy et al., 2021). Several Y-STR and Y-SNP results from contemporary Hungarian-speaking populations have been published (Völgyi et al., 2009; Bíró et al., 2015; Pamjav et al., 2017). However, fewer genetic results exist from populations living in geographically isolated areas in the Carpathian Basin, such as the Bodrogköz (Pamjav et al., 2017).

Based on published uniparental genetic data, the contemporary Hungarians possess some genetic similarities to ancient conquerors in the Carpathian Basin, as well as to Central/Inner Asian populations and populations in the Ural Mountains (Bíró et al., 2015; Fehér et al., 2015; Neparáczki et al., 2017a; Pamjav et al., 2017; Huang et al., 2018; Dudás et al., 2019; Csáky et al., 2020). The core of the Hungarian Conquerors may have originated from Inner Asia/Southern Siberia, and then likely admixed with the peoples they encountered during their migration westwards. Examples of this may be found in the Ural and the Caucasus Mountains. The Y-chromosomal relationship for Inner Asian origin is the presence of haplogroups C-M86, R1a-Z93, and N-M46 in ancient or present-day Hungarian populations (Pamjav et al., 2017; Fóthi et al., 2020). The Y-chromosomal relationship for Ural or Caucasian origin is the presence of Y-haplogroups N-L1034, R1b-L23, Q-M242, and G2a-L156 in Hungarian conquerors or modern Hungarian populations (Bíró et al., 2015; Pamjav et al., 2017; Fóthi et al., 2020). The mtDNA relationship for Inner Asia/Southern Siberia origin is the presence of mtDNA haplogroups A, B, H6, and T1a* in the Hungarian Conquerors or present-day Hungarians (Neparáczki et al., 2017a). The mtDNA haplogroup X2f represents a genetic link to Caucasian populations (Neparáczki et al., 2017a). Other mtDNA studies of early Hungarians revealed a diverse composition of haplogroups with significant Asian affinity (Bogács-Szabó et al., 2005; Branstatter et al., 2007; Tömöry et al., 2007) and concluded that the influence of Asian migration into Eastern Europe is detectable in the Székely (or Sekler) population (Romania) and the relatively high proportion of Asian mtDNA haplogroups A, B, C, G, and Y, totaling 7.9% altogether, may be an indicate an influx of Asian nomads into the Székely population in medieval Transylvania (Branstatter et al., 2007).

Thus, the present-day Hungarian speakers have a very complex genetic history. Moreover, the population consists of several ethnic groups in the Carpathian Basin, such as the Székely, Csángó, Palóc, Jász, and Kun. However, Y-chromosomal studies about them and other regional population groups are still limited.

In our previous study, we studied paternal genetic composition of the Bodrogköz population in Eastern Hungary and found genetic similarities to that of ancient Hungarians (Pamjav et al., 2017). To continue the study, we analyzed two Hungarian speaking populations, the Rétköz population in Eastern Hungary, as well as the Váh valley population in Slovakia. To explore their population histories, all Rétköz and Váh valley samples were surveyed for 23 Y-STRs and over 40 Y-SNPs. The resulting data were compared to our published Y-chromosomal data from Eurasian populations and contemporary Hungarians, as well as to Eurasian populations from other published studies.

The Rétköz is a region in Eastern Hungary, whereas the Váh valley is in Western Slovakia. The populations of both places live within the Hungarian-Slovakian contact zone and are Hungarian-speakers. The Rétköz is near the Bodrogköz (Pamjav et al., 2017), but genetic studies have not been conducted in this region or in the Váh valley. The common features of these three regions (Bodrogköz, Rétköz and Váh valley) are that ancient Hungarian Conquerors lived there and that many of their cemeteries have been excavated.

In this study, we present new Y-STR and Y-SNP data of two Hungarian-speaking populations in the contact zone of Hungarian-Slovakian territories and compare them to contemporary Eurasian and available aDNA data to gain further knowledge about the genetic history of these populations.

2 Materials and methods

2.1 Materials

We collected samples from 106 unrelated males from Rétköz, Hungary and 48 males from the Váh valley, Slovakia (Figure 1). Informed consent was obtained from all live participants included in the study.

2.2 Methods

2.2.1 Testing of Y-STR and Y-SNP markers

Genomic DNA was extracted from buccal swabs using the Investigator kit and EZ1 robotic system (Qiagen, Germany), as described in the manufacturer's instructions. The samples were quantified using the Quantifiler Human kit and the ABI 7500 Real-time PCR System (Thermo Fisher Scientific, Waltham, MA, USA).

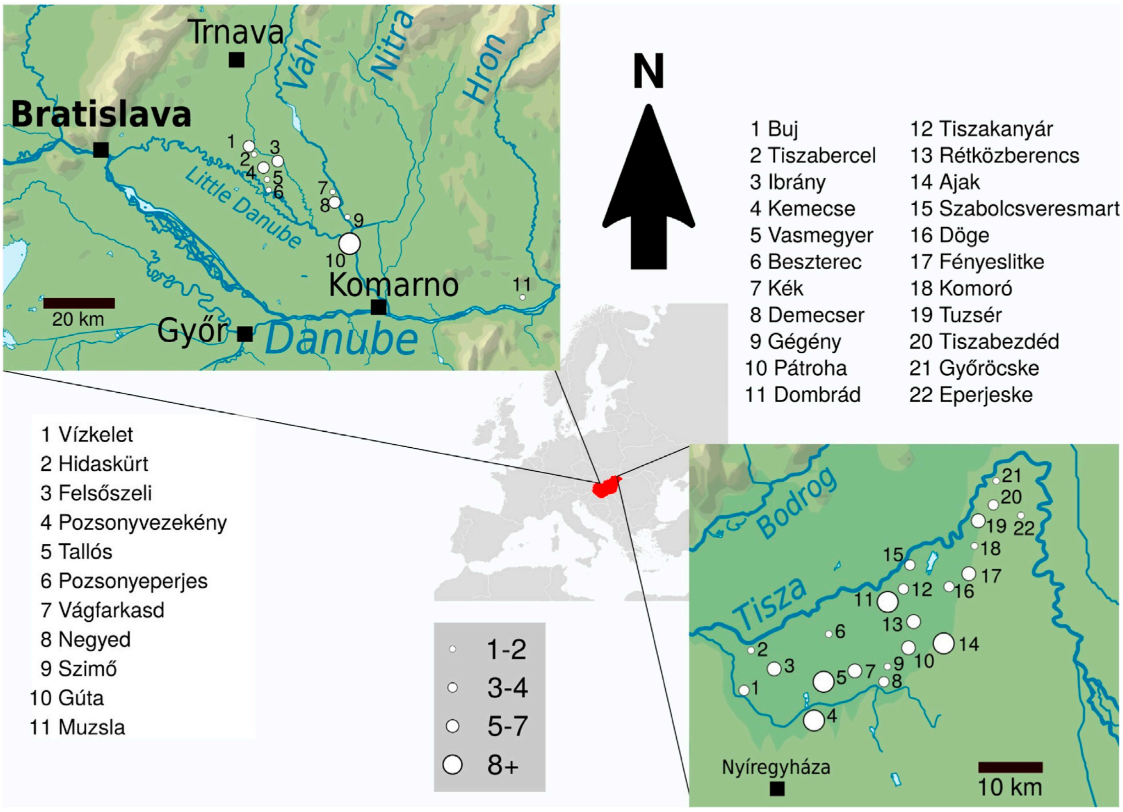


FIGURE 1
The geographical distribution of 48 and 106 non-related Hungarian males from Váh valley and Rétköz, respectively. Left side: Map and list of the sampled Váh valley settlements. Right side: Map and list of the sampled Rétköz settlements. Settlements are marked with circles of different sizes that are proportional to the number of individuals sampled.

DNAfrom the Rétköz and Váh valley populations was surveyed for genetic variation using the Promega PowerPlex Y23 kit. Allele sizing and calling were determined with the ABI3500 Genetic Analyzer and GeneMapper ID-X v.1.4 software. To test for Y-SNP markers (Y chromosomal Single Nucleotide Polymorphism), we performed amplifications of 1–2 ng genomic DNA with Custom TaqMan probes and analyzed the relative fluorescence of the PCR productsin an ABI 7500 Real-time PCR instrument using SDS.1.2.3 software. The SNP markers tested were CT-M168, KT-M9, PR-M45, T-M170, I-M170, I1-M253, I2b-M223, I2a-P37, J-M304, J1-M267, J2-M172, J2-M67, J2b-M12, R-M207, R1-M173, R2-M124, R1a-M198, R1a-SRY1083.1, R1a-M458, R1a-Z93, R1a-Z280, R1b-M343, R1b-P25, R1b-U106, R1b-P312, R1b-M412, R1b-Z2103, D-M174, N-M231, N1-LLy22g, N-L708, N-M46, N-L1034, N-VL29, N-Z1936, N-F4205, R1a-Y2633, and N-Y24365. The haplogroups are described in accordance with the generally accepted nomenclature, as is common practice (Karafet et al., 2008 and the ISOGG).

The Y-STR (Y chromosomal Short Tandem Repeat) haplotypes in this study were sent to the YHRD (accession numbers: YA004754 for the Rétköz and YA004755 for the Váh valley populations).

2.2.2 Phylogenetic analysis

To examine the STR variation within the haplogroups, relational networks were constructed using the Network 10.2.0.0 program (Bandelt et al., 1999). Repeats of the DYS389I locus were subtracted from the DYS389II locus, and the DYS385 locus was excluded because the Network program cannot handle the duplicated locus. To put the results into a more extensive geographical context, we included haplotypes of 10 overlapping evolutionarily stabile loci from other Eurasian populations. The rho statistic in the network program was used to estimate the time to the most recent common ancestor (TMRCA) of haplotypes among the compared haplogroups (Bandelt et al., 1999). Evolutionary time estimates were calculated according to Zhivotovsky et al. and STR mutation rate was assumed to be 6.9 × 10^{-4}/locus/25 years (Zhivotovsky et al., 2004) as is common practice. STR-based TMRCA estimates were for information only and are not discussed in this paper due to their unreliability and insignificance to the primary purpose of the study.

TABLE 1 Haplogroup frequencies of the Rétköz and Váh valley populatios studied.

Haplogroups	Rétköz	%	Váh valley	%	N	%
E1b1-M123	1	0.94	2	4.17	3	1.95
E1b1-M78	7	6.60	7	14.58	14	9.09
G2a-L156	7	6.60	2	4.17	9	5.84
H1a-M82	1	0.94	0	0.00	1	0.65
I1-M253	5	4.70	2	4.17	7	4.55
I2a-P37	7	6.60	8	16.67	15	9.74
I2b-M223	2	1.89	1	2.08	3	1.95
J2a-M67	1	0.94	2	4.17	3	1.95
J2b-M12	3	2.83	0	0.00	3	1.95
J2*-M172	0	0.00	1	2.08	1	0.65
N1c-L1034	1	0.94	0	0.00	1	0.65
N1c-VL29	1	0.94	0	0.00	1	0.65
N1c-Z1936	1	0.94	0	0.00	1	0.65
Q-M242	3	2.83	0	0.00	3	1.95
R1a*-M198	0	0.00	1	2.08	1	0.65
R1a-M458	18	17.00	3	6.25	21	13.64
R1a*-SRY10831	1	0.94	0	0.00	1	0.65
R1a-Z280	21	19.80	8	16.67	29	18.83
R1a-Z93	2	1.89	0	0.00	2	1.30
R1b*-M343/P25	8	7.55	3	6.25	11	7.14
R1b-P312	13	12.30	1	2.08	14	9.09
R1b-U106	3	2.83	7	14.58	10	6.48
22	106	100.00	48	100.00	154	100.00

2.2.3 Genetic structure

Based on the Y-STR haplotypes, pairwise Rst (stepwise mutation) genetic distances were computed with YHRD. org's online AMOVA, and the MDS plot constructed (release 66) (Willuweit and Roewer 2015). Pairwise Fst genetic distances were calculated based on haplogroup frequencies using Arlequin 3.5 software (Excoffier and Lischer, 2010). MDS plot from the Fst values was constructed with the cmdscale function of R (R Core Team, 2021). Rst is an analogy of Fst based on allele size differences; it is defined as the correlation of allele sizes between tested markers within populations. Fst (fixation index) is more efficient when there are high levels of gene flow, whereas Rst reflects population differentiation better under low gene flow (Balloux and Goudet, 2002). That's why we used both Rst and Fst-based metric MDS analyses for population comparisons.

Haplotype and haplogroup frequencies and their diversity values were calculated using the formula from Nei 1973.

3 Results

3.1 Y-chromosome diversity

The haplogroup frequencies of the two populations from the Hungarian-Slovakian contact zone are presented in Table 1. The STR and SNP results of the 106 Rétköz and 48 Váh valley males are shown in Supplementary Table S1. The most frequent haplogroups of the Rétköz population were R1a-Z280 (19.8%), R1a-M458 (17%), R1b-P312 (12.3%), R1b-P25/M343 (7.55%), E1b1-M78 (6.6%), I2a-p37 (6.6%), G2a-L156 (6.6%), and I1-M253 (4.7%). Furthermore, haplogroups J2b-M12, Q-M242, and R1b-U106 accounted for 2.83% of each haplogroup. I2b-M223 was 1.89% of the Rétköz population, whereas the remaining haplogroups, including the Rétköz N-M46 (earlier N1c) chromosomes, which belong to the studied subgroups (N-L1034, N-VL29, N-Z1936), accounted for less than 1% of the lineages.

In the case of the Váh valley males, the most frequent haplogroups were I2a-P37 and R1a-Z280 (both at 16.67%). The frequencies of the remaining haplogroups were as follows: E1b1-M78 (14.58%), R1a-M458 (6.25%), R1b-P25/M343 (6.25%), E1b1-M123 (4.17%), G2a-L156 (4.17%), I1-M253 (4.17%), J2a1-M67 (4.17%), and the remaining haplogroups (2.08%). However, nearly 10% of the present-day Rétköz population may be related to the ancient Hungarians based on haplogroup composition, but the modern Hungarians in the Váh valley lack such relatedness.

The haplotype and haplogroup diversities of the Rétköz group were 1.00 and 0.901, respectively, whereas these values

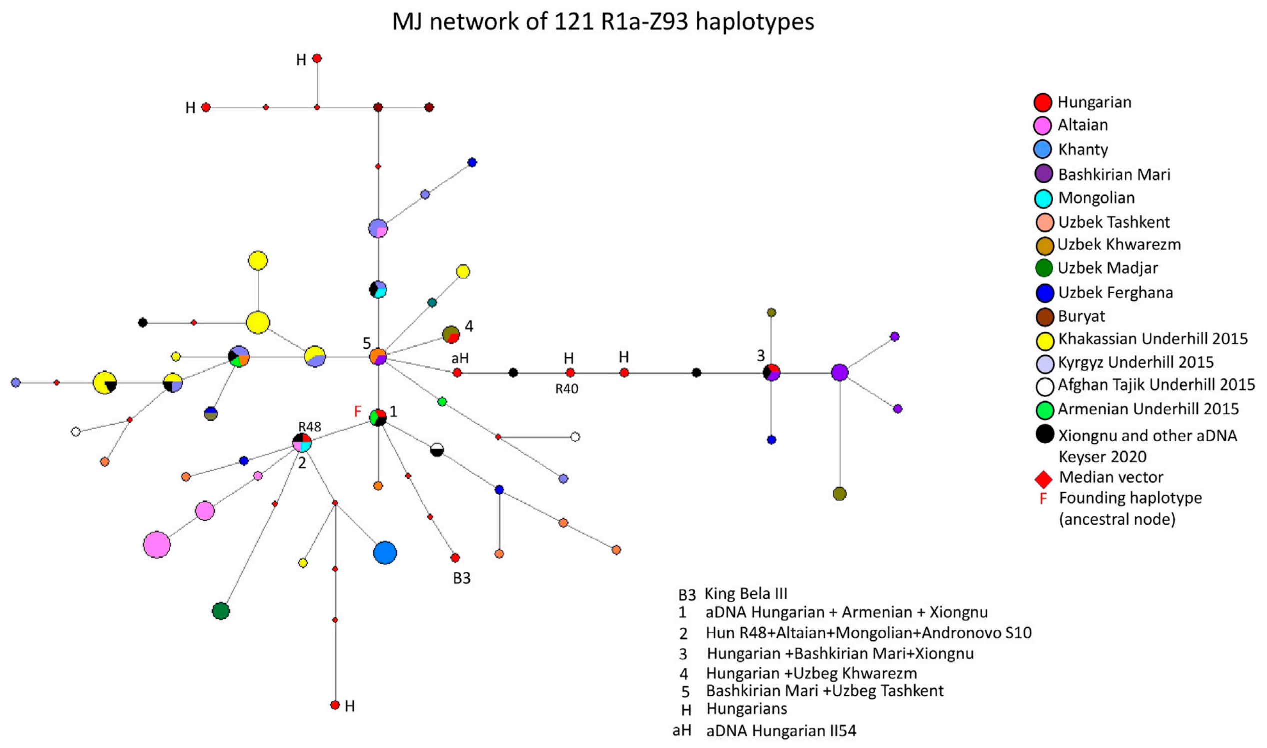

MJ network of 121 R1a-Z93 haplotypes
H
H
5
4
aH
H
R40
H
3
F
1
R48
2
B3
H
Hungarian
Altaian
Khanty
Bashkirian Mari
Mongolian
Uzbek Tashkent
Uzbek Khwarezm
Uzbek Madjar
Uzbek Ferghana
Buryat
Khakassian Underhill 2015
Kyrgyz Underhill 2015
Afghan Tajik Underhill 2015
Armenian Underhill 2015
Xiongnu and other aDNA Keyser 2020
Median vector
F Founding haplotype (ancestral node)
B3 King Bela III
1 aDNA Hungarian + Armenian + Xiongnu
2 Hun R48+Altaian+Mongolian+Andronovo S10
3 Hungarian +Bashkirian Mari+Xiongnu
4 Hungarian +Uzbeg Khwarezm
5 Bashkirian Mari +Uzbeg Tashkent
H Hungarians
aH aDNA Hungarian II54

FIGURE 2
The Median-Joining Networks (MJ) of 121 R1a-Z93 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual.

for the Váh valley were 0.998 and 0.904, respectively. The results show that, in both populations, haplotypes are more diverse than haplogroups.

3.2 Phylogenetic analysis

Based on the Y-STR haplotypes, networks were constructed for haplogroups that may be linked to those of the Hungarian Conquerors, because these haplogroups can be found in ancient Hungarians and originated from Inner/Central Asia. We also included 10 Y-STR haplotypes from other Eurasian populations and aDNA results from published sources to widen the geographical context. We have included six networks (R1a-Z93, N-Tat, Q-M242, R1b-P25/M343, E1b1-M78, and G2a-L156) that are potentially helpful in uncovering the genetic legacy of the populations being studied. Other networks, which were constructed for haplogroups of European origin (R1a-M458, R1a-Z280, R1b-P312, R1b-U106 and I2a-P37) and could not be related to Hungarian Conquerors, were not included in the study. Supplementary Table S2 summarizes data for all the published aDNA samples used in the study including haplotypes, haplogroups, ages and geographic origins.

3.2.1 Median joining network of 121 R1a-Z93 haplotypes

Figure 2 depicts an MJ network of 121 R1a-Z93 haplotypes from the 15 populations tested by us for this study or previously published (Bíró et al., 2015; Underhill et al., 2015; Dudás et al., 2019) and aDNA samples (Olasz et al., 2018; Fóthi et al., 2020; Keyser et al., 2021). Three modern Hungarian and one aDNA samples (II54 from the Hungarian Royal Basilica of Székesfehérvár), including 1 Rétköz sample (Rétköz 40), formed a common branch with three ancient Xiongnu samples (TUK45, TUK04, TUK25) on the right side of Figure 2. On this branch, one modern Hungarian sample shared two haplotypes (1 TUK25 and one Bashkirian Mari) in haplotype cluster 3. Two Uzbek samples (1 Khwarizm and one Fergana sample) can be derived from cluster 3 (see Figure 2). The five Bashkirian Mari and two Uzbek (Khwarizm) samples are clustered one molecular step from cluster three at the end of this branch. Cluster five includes one Bashkirian Mari and two Uzbek males from Tashkent in Uzbekistan and located at one molecular step from cluster 1. Cluster one includes three haplotypes: one Hungarian aDNA (Nagykörös Gr2), one Xiongnu aDNA (TUK09A) and one modern Armenian sample, separated by one molecular step (DYS389I) from cluster 5. The founding haplotype, which may have arisen in the common ancestor of

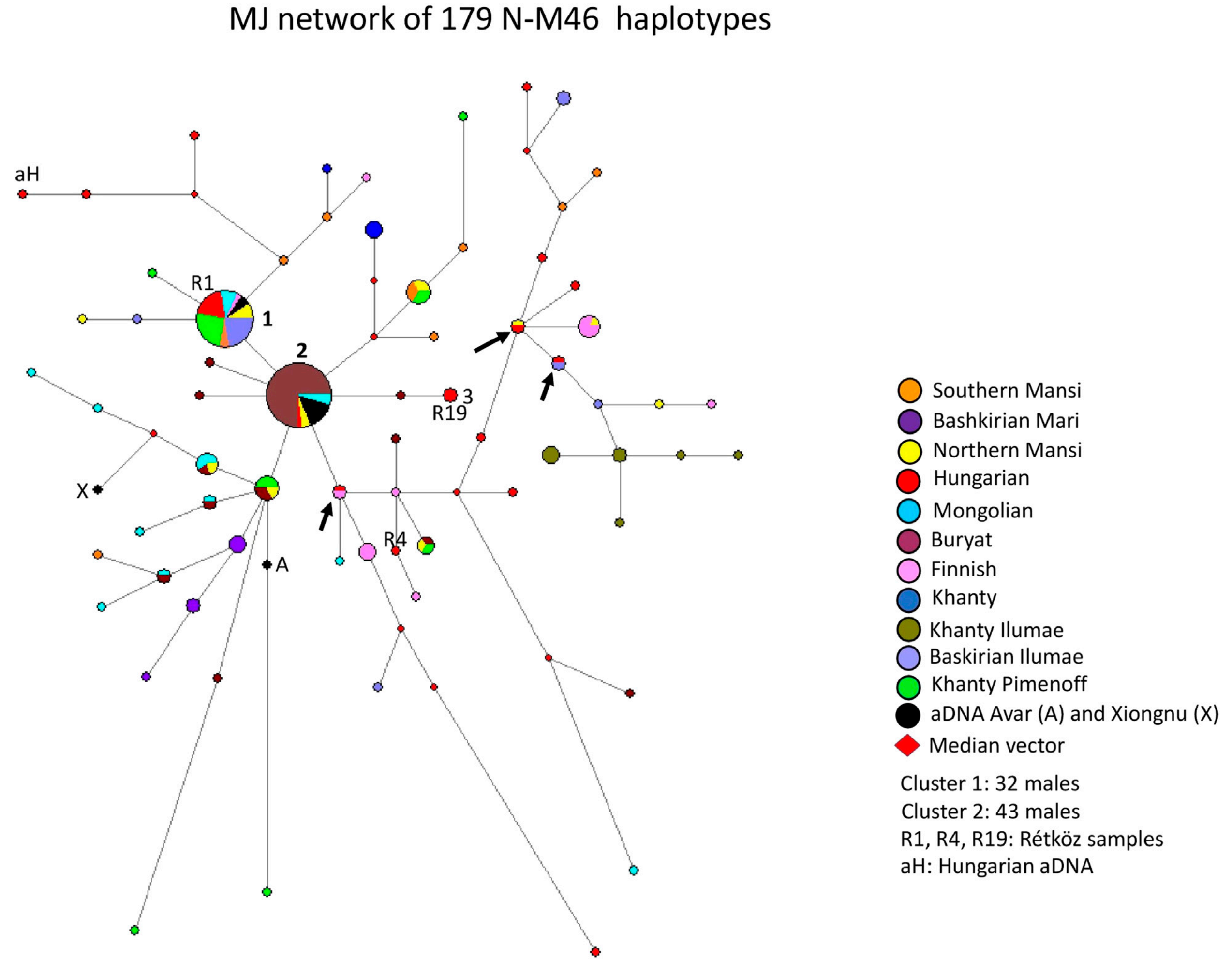


FIGURE 3
The Median-Joining Networks (MJ) of 179 N-M46 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual.

populations or males and is shared by them. Based on this, cluster 1 may be the founding haplotype, as it contains two ancient haplotypes from males that lived 1,000 to 2,500 years ago. All the samples and branches are derived from this haplotype. Cluster two includes one Rétköz Hungarian (Rétköz = R48), one Mongolian, one Altaian, and one Andronovo aDNA (S10) samples. Cluster four can be separated one molecular step from cluster five and includes one Hungarian and two Uzbek males from Tashkent in Uzbekistan. The Hungarian King Bela III is found three molecular steps away from cluster 1 (see B3 in Figure 2). The remaining three Hungarian haplotypes are outliers in the network and are not shared by any of the samples.

The other studied samples in the network either form independent clusters, such as Altaians, Khakassians, Khanties, and Uzbek Madjars, or are scattered within the network. Other aDNA samples (S26, ARZT1, ARZT28, and MN376 from Keyser et al., 2021) in the network form a different branch, with almost all Khakassian samples located on the left side of Figure 2. The age of accumulated STR variation (TMRCA = Time to Most Recent Common Ancestor) within the R1a-Z93* lineage for 121 samples is estimated as 13 ± 3 kya (95% CI = 10.0–16.0 kya), considering that cluster one is the founder haplotype, which is higher than that of the SNP-based calculation (4.6 kya, 95% CI: 4.2–5.0 kya) (Adamov et al., 2015; www.yfull.com).

3.2.2 Median-joining network of 179 N-M46 haplotypes

A median-joining network of 179 N-M46 (previously N1c, Karafet et al., 2008) haplotypes was generated using populations we previously studied (Bíró et al., 2015; Fehér et al., 2015), as well as those researched by others (Pimenoff et al., 2008; Ilumäe et al.,

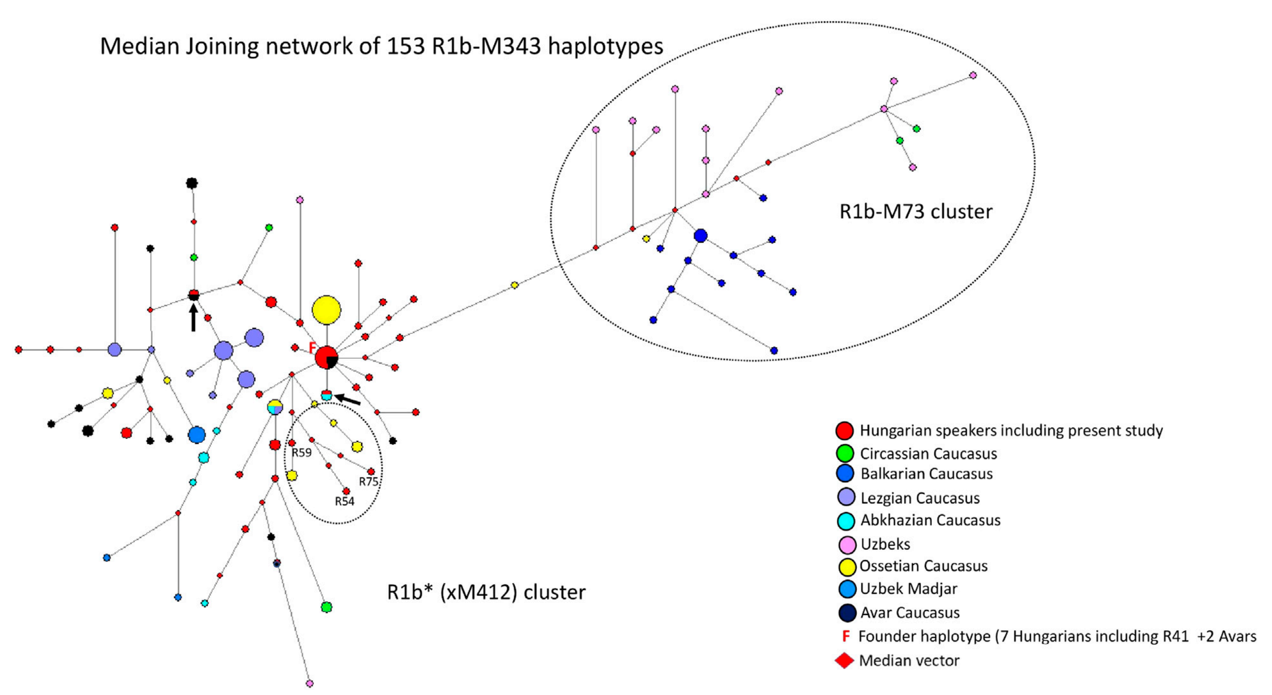


FIGURE 4
The Median-Joining Networks (MJ) of 153 R1b-M343* (P25) haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual.

2016) (Figure 3). The founder N-M46 (Tat = M46) haplotype was shared by 32 samples from eight populations (8 Khanty from Pimenoff et al., 2008; three Northern Mansi; three Mongolian; six Hungarian-speakers, including the Rétköz 01 sample; two Southern Mansi; seven Bashkirian from Ilumäe et al., 2016; two ancient Avars from Csáky et al., 2020, and one Finnish), as seen in Figure 3 (haplotype cluster 1). Cluster one is the founding haplotype. Cluster 2, another large cluster, includes 43 haplotypes (32 Buryat; two Mongolian; one Hungarian; two Northern Mansi; and six ancient Avars from Csáky et al., 2020) from five populations, as shown in Figure 3. Cluster two is located one molecular step from cluster 1. The only difference is in the DYS391 locus, with allele 11 in cluster one and allele 10 in cluster 2. As seen in Figure 3, 73% of Buryats belonged to cluster 2, indicating that the Buryats we studied belong to a young and isolated population (Dudás et al., 2019). Two Hungarian males, including Rétköz 19 (R19) sample, derives from cluster two *via* one Buryat haplotype (see three in Figure 3). One Hungarian aDNA haplotype (Ö52/50, Fóthi et al., 2020) was positioned six mutational steps away from cluster one and formed a haplotype branch with two present-day Hungarian males (see aH in Figure 3). Three modern Hungarian males shared 1-1 haplotypes with Finnish, Northern Mansi, or Bashkirian males (see black arrows in Figure 3).

The other population samples included in the network formed independent clusters, such as Bashkirian Mari, Finns, Khanties, Bashkirians, and Khanties from another research (Ilumäe et al. (2016). They also shared haplotype clusters or were scattered in the network.

The age of accumulated STR variation (TMRCA) within N-M46 lineage for 179 samples is estimated as 11.3 ± 3.3 kya (95% CI = 8.0–14.6 kya), considering cluster one is the founder haplotype, which is similar to its sequence-based calculation of 13 kya (95% CI:11.3–14.6 kya) (Ilumäe et al., 2016).

3.2.3 Median-joining network of 153 R1b-M343 haplotypes

A median-joining network of 153 R1b-M343/P25 (Eastern R1b-L23) haplotypes was generated using samples from FTDNA, populations we previously studied (Balanovsky et al., 2011; Bíró et al., 2015) (Figure 4). The founder R1b*-M343 (L23) haplotype was shared by nine samples, including the Rétköz 41 (R41) sample, from two populations (7 Hungarians and two Avars from the Caucasus), as shown in Figure 4. The majority Hungarian haplotypes, including four Rétköz (R28, R52, R39 and R37) and one Váh valley samples (V24), appear to be descended from the founding haplotype (F = founder), because these haplotypes differ one molecular step from the founding haplotype. The pattern of these haplotype clusters is starlike,

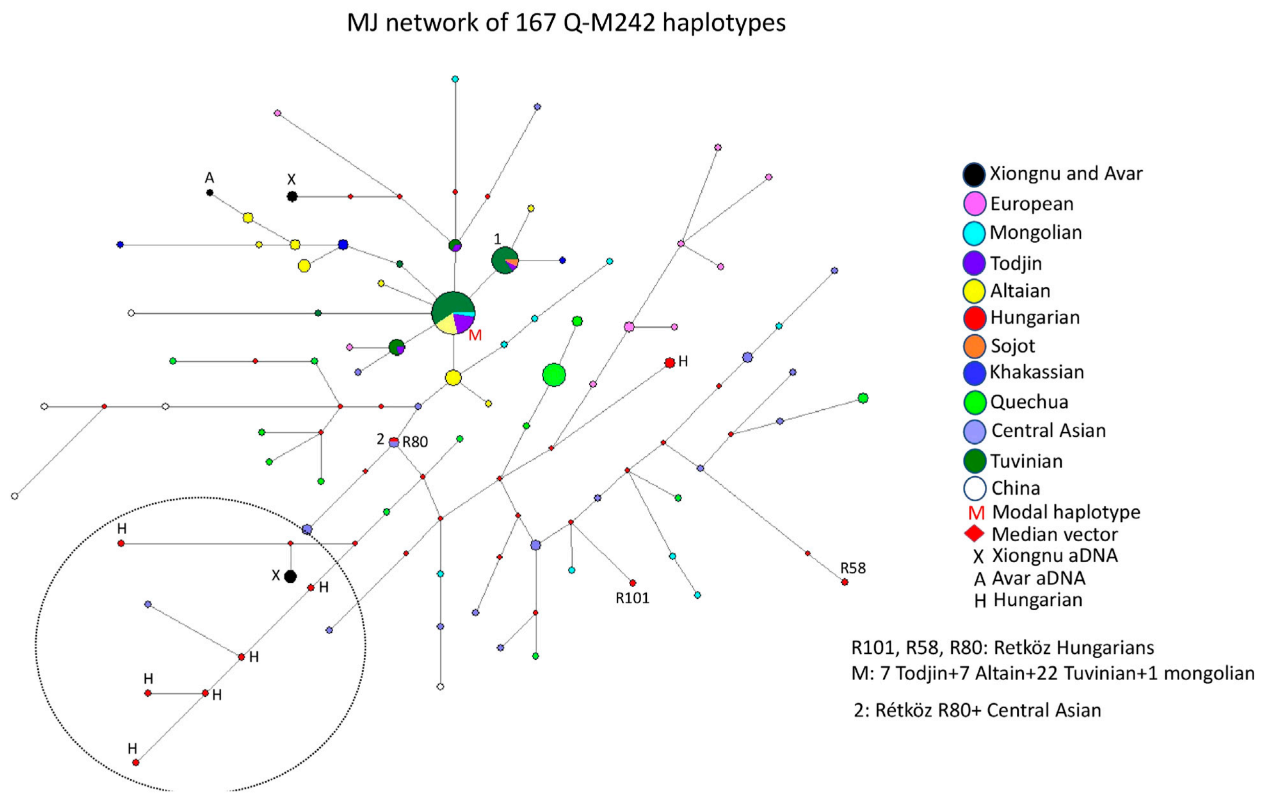


FIGURE 5
The Median-Joining Networks (MJ) of 167 Q-M242 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual.

representing a set of closely related haplotypes of Hungarian males. Three Rétköz R1b samples (R54, R59 and R75) clustered with Ossetian samples from the Caucasus (indicated by the red circles in a dashed circle under the founding haplotype cluster in Figure 4.). In addition to the founding haplotype, some Hungarian haplotypes were shared with Abkhazian (turquoise and red circles) or with Hungarians and Avars from the Caucasus (red and black circles), which are marked with an arrow. All other Hungarian haplotypes were scattered within the network. It is interesting that either Hungarian haplotypes shared a common haplotype with Avar males or clustered with Avar haplotypes in the network (red and black circles).

The other population samples in the network formed independent clusters, such as Lezgian, Ossetian, and Circassian from the Caucasus or were scattered in the network. Primarily Balkarian (Caucasus) and Uzbek males form an independent branch (R1b-M73) on the right side of the network (Figure 4). The age of accumulated STR variation within the R1b-M343/P25 lineage for 153 samples is estimated to be 22.8 ± 3.8 kya (95% CI = 19.0–26.6 kya), considering the founder haplotype is the ancestral one (F in Figure 4), which is in agreement with the time calculated on sequence data (yfull.com). For the R1b (xM412) branch, the age of accumulated STR variation is 17.7 ± 5.2 kya (95% CI = 12.5–22.9 kya), which is much higher than whtat Myres et al., 2011 calculated.

3.2.4 Median-joining network of 167 Q-M242 haplotypes

The median-joining network of 167 Q-M242 haplotypes from 12 populations does not show a star-like pattern, but rather a more diverse one. However, it was simple to analyze the genetic relationship of the Hungarian samples (Figure 5). The largest haplotype cluster, to which a relatively large number of individuals belong, is shared by four populations (see M on Figure 5). This cluster includes 22 Tuvinian, seven Todjin, seven Altaian, and one Mongolian haplotypes. Another large haplotype cluster (cluster 1) includes 14 males from three populations (2 Sojot, one Tojin, and 11 Tuvinian males), whereas almost all other haplotypes are scattered within the network.

The Rétköz 80 (R80) sample shares the same haplotype as an Uzbek male, and the Rétköz 101 (R101) sample clusters with a Mongolian male. The third Rétköz male (R58) is located six steps away from an Uzbek sample and is an outlier.

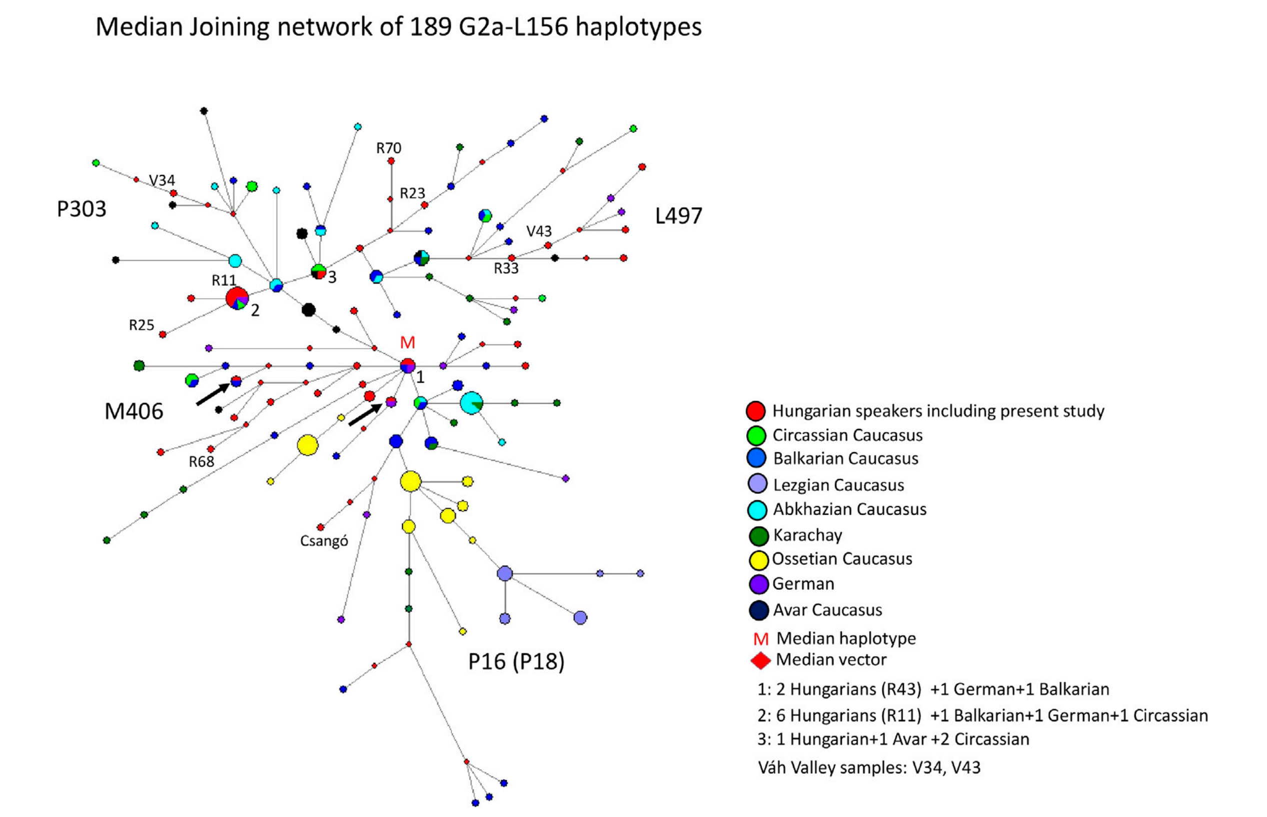

Median Joining network of 189 G2a-L156 haplotypes
P303
V34
R70
R23
L497
V43
R33
R11
3
2
R25
M
1
M406
R68
Csangó
P16 (P18)
Hungarian speakers including present study
Circassian Caucasus
Balkarian Caucasus
Lezgian Caucasus
Abkhazian Caucasus
Karachay
Ossetian Caucasus
German
Avar Caucasus
M Median haplotype
Median vector
1: 2 Hungarians (R43) +1 German+1 Balkarian
2: 6 Hungarians (R11) +1 Balkarian+1 German+1 Circassian
3: 1 Hungarian+1 Avar +2 Circassian
Váh Valley samples: V34, V43

FIGURE 6
The Median-Joining Networks (MJ) of 189 G2a-L156 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual.

It is noteworthy that some Hungarian samples, primarily Hungarian-speaking Csángó and Székely males in Transylvania, Romania form a separate branch with the ancient Xiongnu males (TUK01, TUK20, and TUK26) (see the black and red circles within the dashed circle in the lower left part of Figure 5), indicating a close genetic affinity with each other. Two ancient Xiongnu (TUK43 and TUK44) samples and one ancient Avar sample from published sources (Csáky et al., 2020; Keyser et al., 2021) are more closely related to a European (Lithuanian) and an Altai sample (see the top of Figure 5.), which indicate different histories. We constructed another network, including Jewish and Balkarian samples, with 192 haplotypes. Two Hungarian males (Hu423 and E699) formed a cluster with the Jewish samples, suggesting a genetic link (data not shown). These Hungarian samples form a cluster with European samples on the right side of Figure 5 (see red H and pink circles). Three Hungarian samples (Csangó 68, Székely 180, and Hu1769) form a branch with Balkarian males (data not shown). These three Hungarian samples form a branch with other Hungarian and Xiongnu samples in Figure 5. Based on this network, we have distinguished the Hungarian samples that have a genetic relationship with the Jewish and Balkarian samples.

The age of accumulated STR variation within the Q-M242 lineage for 167 samples is estimated to be 17.7 ± 3.6 kya (95% CI = 14.1–21.3 kya), considering that the modal haplotype (Figure 5) is the founder, which agrees with the previous calculated 15–25 ky (Karafet et al. (2008).

3.2.5 Median-joining network of 189 G2a-L156 haplotypes

The MJ network of 189 G2a-L156 haplotypes is depicted in Figure 6. The samples belong to four subgroups (P303, L497, M406 and P16). The model haplotype cluster (cluster 1 = M in Figure 6) is shared by three populations, including two Hungarian, one German and one Balkarian males. The haplotype cluster is likely to be the founder haplotype, as all branches are derived from this and is the median center of the branches. The biggest haplotype cluster is shared by four populations, including six Hungarians (Rétköz 11 = R11), one German, one Balkarian, and one Circassian chromosome fromthe Caucasus (cluster two in Figure 6). Cluster three is shared by three populations, including one Hungarian, one Avar and two Circassian males from the Caucasus. One Hungarian male matches a German haplotype (red-purple

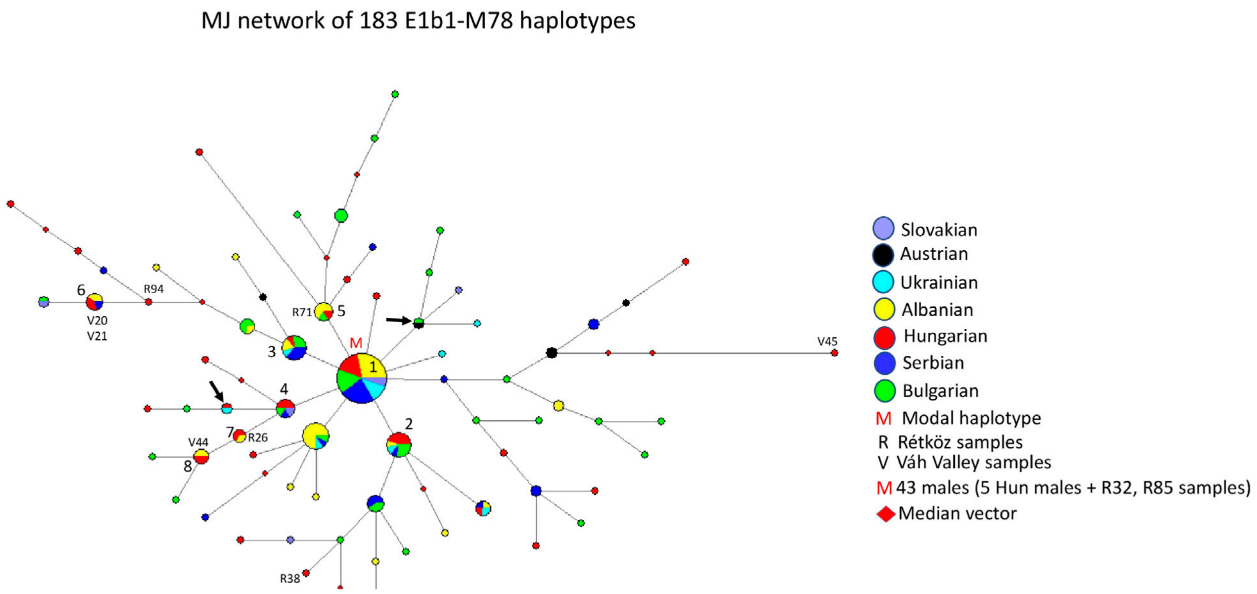


FIGURE 7
The Median-Joining Networks (MJ) of 183 E1b1-M78 haplotypes. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual.

circle), and another Hungarian male matches with a Balkarian male (red-dark blue circle in Figure 6), indicating a common origin with Hungarians. We marked these haplotypes with an arrow in Figure 6, as well. All other Hungarian males derive from clusters 1-3, and a shared haplotype was not observed. It appears that the Hungarian haplotypes, including Rétköz (R11, R23, R33, R68 and R70) and Váh valley (V34, V43) samples, are generally clustered together on a common branch with Balkarian/Karachay, Avar, or German haplotypes, suggesting that these chromosomes have common origin. Ossetian, Lezgian, and Abkhazian males from the Caucasus form independent clusters, except for the Circassians, which are scattered within the network.

The age of accumulated STR variation within the G2a1-L156 lineage for 189 samples is estimated to be 19.2 ± 3.7 kya (95% CI = 15.5–22.9 kya), considering the modal haplotype (Figure 6) is the founder one. This is in line with the 15.0 ky calculated by Rootsi et al. (2012).

3.2.6 Median-joining network of 183 E1b1-M78 haplotypes

Figure 7 depicts the median-joining network of 183 E1b1-M78 haplotypes. The network shows a star-like pattern. The biggest cluster (cluster 1 = M) is the modal haplotype shared by six populations, including 43 males. The model haplotype includes seven Hungarian (R32 and R85), 12 Albanian, five Ukrainian, 10 Serbian, seven Bulgarian, and two Slavic haplotypes. There are several haplotype clusters showing different admixture of the seven populations researched in the study, such as cluster 2 (5 Hungarian, one Albanian, one Ukrainian, one Serbian, and three Bulgarian males), cluster 3 (1 Hungarian, two Albanian, one Ukrainian, four Serbian, and three Bulgarian males), cluster 4 (3 Hungarian, one Bulgarian, one Serbian, and one Slavic males), cluster 5 (4 Albanian, one Rétköz Hungarian, and one Bulgarian males), cluster 6 (2 Váh valley Hungarian, two Albanian, and one Serbian males), cluster 7 (1 Hungarian, one Rétköz and one Albanian males), and cluster 8 (1 Hungarian, one Váh valley, and two Albanian males). The remaining Hungarian and other samples are scattered within the network or have shared haplotypes, like Ukrainian-Hungarian or Austrian-Bulgarian (see arrows in Figure 7).

The age of accumulated STR variation within the E1b1-M78 lineage for 183 samples is estimated to be 6.7 ± 1.3 kya (95% CI = 5.4–8.0 kya), considering the modal haplotype (see M in Figure 7) is the founder, which is in line with the arrival of Neolithic farmers in Europe (Battaglia et al., 2009).

3.3 Genetic structure

We constructed a non-metric multidimensional scaling (MDS) based on Y-chromosomal haplotypes (2,405 haplotypes) that consisted of 23 STR loci available from 14 populations (www.yhrd.org). The Rst-genetic distances and Rst *p*-values of the

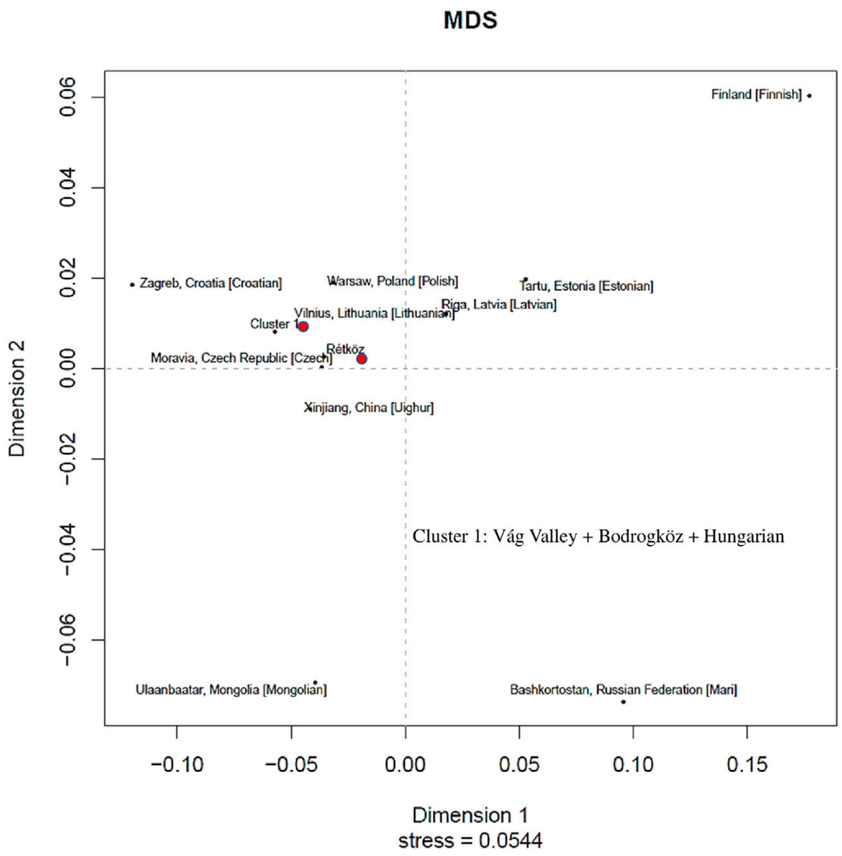


FIGURE 8
Multidimensional scaling (MDS) plot constructed on Rst genetic distances of 10 STR-based 2405 Y haplotypes of 14 populations compared (www.yhrd.org). The Rst-genetic distances and Rst *p*-values of the studied populations are presented Supplementary Table S2.

studied populations are presented in Supplementary Table S3. As shown in Figure 8, the Rétköz population had close Rst genetic distances (<0.05) with the Czech, Estonian, Xinjiang Uighur in China, Latvian, Lithuanian, Polish and cluster one populations. We used the relaxing MDS calculation to cluster the populations and, in this case the Rst threshold value of the indistinguishable populations was 0.01. As a result, Váh valley, Bodrogköz, and Hungarian populations formed by cluster 1. Based on the Rst genetic distances, the most distant populations with the Rétköz population were the Finnish (0.2593) and Bashkirian Mari (0.1375).

The cluster one populations showed close genetic affinities (<0.05) with most of the European populations that were compared, such as Croatian, Czech, and Polish. Interestingly, cluster one population is closely related to the Xinjiang Uighur population in China. Finnish (0.2844), Bashkirian Mari (0.1722), and Estonian (0.1181) populations, however, are genetically farthest from cluster 1.

The haplogroup frequency data used for population comparisons, as well as their corresponding references are included in Supplementary Table S4. Pairwise Fst-distances and *p*-values for 61 populations, including Rétköz, Váh valley, and other Eurasian populations from published sources were calculated as shown in Supplementary Table S5 and presented in a metric MDS plot (Figure 9). Between the two studied populations, the pairwise Fst-distance was insignificant ($p > 0.05$) (Supplementary Table S5). Furthermore, Rétköz and Váh valley had insignificant Fst-values, with 5 and 11 populations, respectively. Among them, three overlapped: two Hungarian (Bodrogköz and Szeged) and a Slovenian population. Czech and Western Slovak have an insignificant Fst-distance from Rétköz, but not from Váh valley. Váh valley showed an insignificant Fst-distance from eight other populations: Bulgarian, Gagauz, Greek, Moldavian, Serbian, and three Hungarian (Csángó, Székely, and a representative Hungarian population).

The location of the studied populations on the MDS plot is consistent with the geographical distances between them (Figure 9). Populations from the same geographic region were clustered together. Hungarian populations, however, overlapped with Balkanian populations. The Rétköz and Váh valley samples were shown to be relatively far from the Hungarian Conqueror samples.

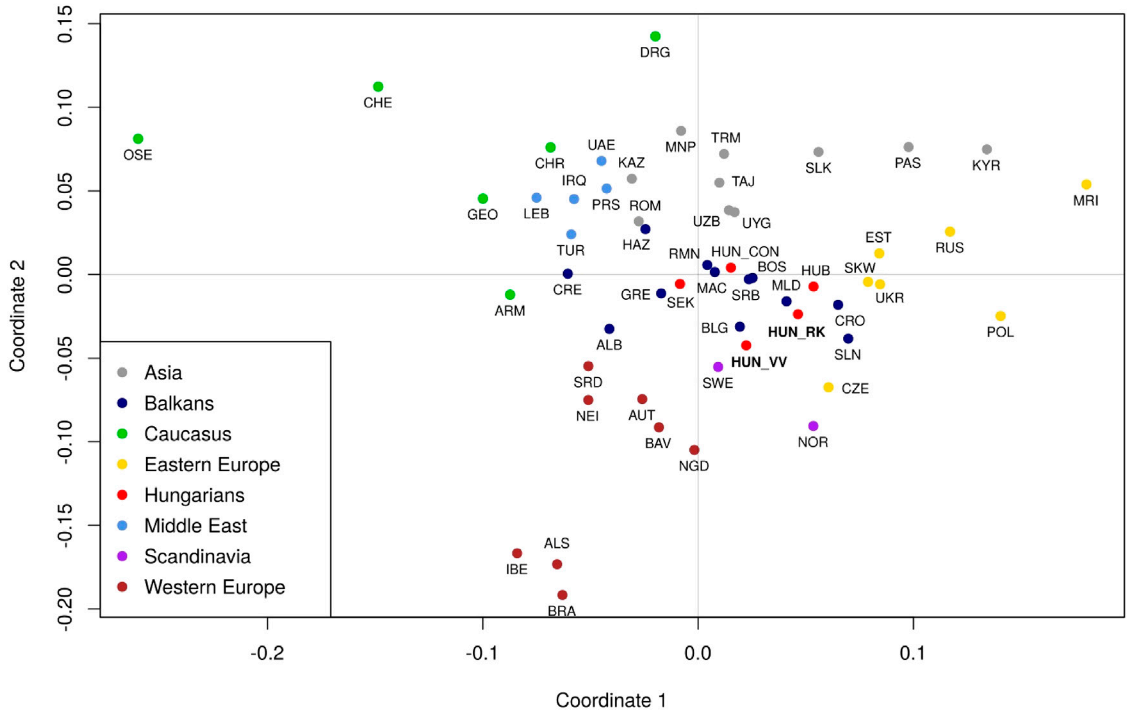


FIGURE 9
Multidimensional scaling (MDS) plot based on pairwise Fst genetic distances of 61 populations. The populations are colored based on their geographical locations. The red dots indicate Hungarian populations. Three populations are highlighted: Váh valley (HUN_VV) and Rétköz (HUN_RK) in bolded red font and Hungarian Conquerors (HUN_CON) in italic red font. Population abbreviations are defined in Supplementary Table S3. Pairwise Fst genetic distances and p-values between 61 populations were calculated as shown in Supplementary Table S4.

A notable difference between the two populations is that the Rétköz population was genetically nearer to the Eastern European populations, whereas the Váh valley population was genetically closer to Western European populations. These observations are consistent with the populations' geographical locations.

4 Discussion

4.1 Phylogenetic analysis

The primary objectives of this study were to create a male phylogeny of the Hungarian-speakers in Rétköz and the Váh valley and to compare them with other Eurasian populations, with aDNA samples from their probable ancestral geographical origin, and with previously studied populations.

To examine the genetic variation within the Hungarian groups, we used evolutionarily stable binary markers (SNPs) to define the haplogroup of each Y-chromosome, and then examined the STR-defined variation within each haplogroup. The established geographical specificity of Y haplogroups means that the haplogroups observed in the Hungarian-speakers may be attributed to their geographical origins or may be related to the Hungarian Conquerors. These conclusions, however, remain uncertain.

The overall pattern of haplogroup distributions in the studied populations was similar, but haplogroups R1a-Z93, N1c-L1034, N1c-VL29, N1c-Z1936, and Q-M242 appeared only in the Rétköz population (Table 1). Haplogroups G2a-L156 and R1b-M343/P25 (L23) were observed more frequently in the Rétköz population. Both lineages range predominantly across the Caucasus and Asia. As such we focused on the genetic history of these haplogroups.

4.1.1 Median-joining network of 121 R1a-Z93 haplotypes

Underhill et al. state that the paragroup R1a-Z93 was most common in the South Siberian Altai region but that it also occurred in Kyrgyzstan and, in all Iranian populations. R1a-Z2125 occurred at highest frequencies in Kyrgyzstan and among Afghan Pashtuns. Additionally, the R1a-Z93 haplogroup was also common in Afghan (Tajik) and Caucasian ethnic groups (Tajik). As such, we included the populations (Z93 haplotypes) from the regions analyzed by other researchers (Underhill et al., 2015), ancient Hungarian, Xiongnu, and Avar samples (Olasz et al., 2018; Csáky et al., 2020; Fóthi et al., 2020; Keyser et al., 2021) in the network analysis.

It is surprising that the present-day Hungarian, ancient Hungarian and Xiongnu samples are located adjacent to each other on this branch, indicating that 1-1 mutation step separates them. Keyser et al., 2021 demonstrated that ancient Xiongnu (TUK45, TUK04, TUK25 and TUK09A) samples belonged to haplogroups R1a-Z93 (Z2125), which are also included in our network. These samples clustered with our Hungarian aDNA (II54) and three present-day Hungarian samples, including Rétköz 40 (R40), in the branch mentioned above (see Figure 2). Furthermore, a Hungarian aDNA haplotype (Nagykörös Gr2) shared the same haplotype as the Xiongnu TUK09A sample (cluster one in Figure 2.). Likewise, a modern Hungarian sample shared the same haplotype as the Xiongnu TUK25 sample (cluster three in Figure 2).

aDNA studies showed that the Hungarian King Béla III, and another sample (II54) from Royal Basillica belonged to haplogroup R1a-Z93 (Olasz et al., 2018), and two R1a-Z93 samples were found among Hungarian Conqueror population as well (Fóthi et al., 2020). Additionally, some Xiongnu R1a-Z93 haplotypes matched those of the Hungarian Conquerors (Keyser et al., 2021). These observations confirm a genetic relationship between Hungarian Conqueror and the ancient Xiongnu people, suggesting descent from a shared common ancestor. These observations also reflect that the descendants of the Xiongnu people and the ancient Hungarians are found among modern Hungarian populations, indicating the presence of a common genetic trace (i.e., genetic continuity).

4.1.2 Median-joining network of 153 N-M46 haplotypes

The Y-chromosomal haplogroup N-M231 is widespread from Scandinavia to the Kamchatka in North Eurasia and is the most frequent single haplogroup in Siberia (Karafet et al., 2008; Rootsi et al., 2012). Based on the geographic distribution of the most basal N*-M231, its lineage most likely originated in Southeast Asia, whereas its most widespread subgroup is N-M46 (Rootsi et al., 2012). Other authors consider Southern Siberia as its geographical origin (Derenko et al., 2007).

The Hungarian Y-chromosomal gene pool has only a small percentage of N-M46 (1%; Tat = M46) and has a distribution typical of East-Central Europe (Völgyi et al., 2009). However, its incidence is higher among the Hungarian-speaking Bodrogköz (6.2%) and Székely (6.3%) (Transylvania, Romania) populations (Biró et al., 2015; Pamjav et al., 2017). Based on Hungarian aDNA studies, N-M46 is detected at a higher frequency in Hungarian Conquerors (Neparaczki et al., 2019; Fóthi et al., 2020). These results support genetic continuity between the ancient Hungarian and the present-day Hungarian populations.

The Rétköz N-M46 sample 1 (R1) from this study and four other Hungarian samples belong to haplotype cluster one and share 31 haplotypes with 8 Eurasian populations (Hungarian, Mongolian, Southern Mansi, Avar, Northern Mansi, Finnish, Bashkirian, and Khanty), indicating they may all share a common genetic history (Figure 3). The second Rétköz (R19) N-M46 sample shared a haplotype with the general Hungarian population (see three in Figure 3.). The third Rétköz (R4) haplotype does not share a haplotype and instead clusters with Finnish haplotypes. This is consistent with the N-Z1936 subgroup, as this SNP is common in the Finnish population. Other modern Hungarian N-M46 samples, including Hungarian aDNA (aH: Őrkút Gr50), are located on the top or the right side of the network, along with Bashkirian, Finnish, and samples originating in the Ural Region (Southern and Northern Mansi, Khanty). Three Hungarian males share a haplotype with the Finnish, Northern Mansi, and Bashkirian samples, suggesting a shared ancestor (see the black arrows). One aDNA Xiongnu (X: TUK30A) sample from Mongolia clusters with Mongolian samples on the left side of the network, indicating a genetic relationship between these groups.

4.1.3 Median-joining network of 153 R1b-M343 haplotypes

The most frequent Western European lineage, haplogroup R1b-M269, was originally thought to have originated in the Palaeolithic. Recent analysis, however, suggests a Neolithic origin (Batini et al., 2017). Most R1b-M412 chromosomes belong to Western European males, but another subgroup, R1b-L23, is commonly referred to as "Eastern European R1b". Its frequency among Turkish, Caucasian, and some SE European and Circum-Uralic populations is about 10% (Myres et al., 2011).

Our network analysis showed that the R1b*-M343 (xM412) haplotypes are divided into two subclades: R1b-L23 (xM412) and R1b-M73. The Rétköz and other Hungarian haplotypes cluster with the L23 samples included in the study and share haplotypes with Hungarian, Abkhazian, and Avar males from the Caucasus, suggesting a genetic link between them (see Figure 4). These Avars are unlikely the same as the Pannonian Avars in the Carpathian Basin because present-day examples live in the Northeastern Caucasus, primarily in Dagestan, Kalmykia, and Chechnya in Russia. The earliest mention of the Avars in European history is by Priscus, who reported in 463 C E that a joint delegation of Saragurs, Urogs, and Unogurs requested an alliance with Byzantium. The delegation claimed that in 461 CE, their peoples were displaced by the Sabirs due to pressure from the Avars (Priscus, 463 CE). The German researcher Karl Heinrich Menges stated that these Avars have nothing to do linguistically with the Proto-Mongolian Avars of the Great Migrations, but as some of the latter may have taken refuge in this region, their name has become the common name of the Avars of present-day Dagestan, who have no name of their own (Menges, 2011).

The haplotype analysis revealed a genetic relationship between the contemporary Hungarian-speakers and Caucasian populations (Avars, Abkhazians, and Ossetians) included in this study. It should be noted, however, that the peoples of East-Central Europe received Central-Inner Asian and Caucasian genes before and after the ancient Hungarians settled in the Carpathian Basin (pre-Hungarians: Sarmatian-Alans, Huns, Avars, and Onogur-Bulgars; post-Hungarians: Pechenegs, Jassic peoples, and Cumans) (Lipták 1979). Based on aDNA studies, haplogroup R1b-L23 was found in Hungarian Conquerors (Neparáczki et al., 2019; Fóthi et al., 2020), supporting the genetic footprint observed in our results, but it was not possible to determine when and where this genetic trace was introduced into the gene pool of the ancient Hungarians.

4.1.4 Median -joining network of 167 Q-M242 haplotypes

The human Y-chromosome haplogroup Q-M242 likely originated in Central Asia and Southern Siberia 15–25 kya (Karafet et al., 2008), and then subsequently diffused eastward, westward, and southward (Di et al., 2013; Huang et al., 2018). Haplogroup Q-M242 reaches its highest frequencies in Siberia, particularly in Kets (90–94%) and Selkups (66–71%) but is rarely found in Western, Southern, and Southeastern Asia (Di et al., 2013; Huang et al., 2018). Subclade Q-M120 occurs in Eastern Asia and migrated from north to south with ancestors of the Han Chinese during the Neolithic period (Gayden et al., 2007; Zhao et al., 2015). Subclades Q-M25 and Q-M346 spread widely in Eurasia. Q-M25 reaches its highest frequency among the Turkmen (34–43%) and spread from Central Asia into Western Asia and Hungary, whereas Q-M346 appears in most parts of Eurasia (e.g., Central, Western, and Southern Asia), as well as the Comoros Islands of Africa (Sengupta et al., 2006; Malyarchuk et al., 2011; Huang et al., 2018). Subclade Q-M3 is present only in Indigenous Americans (Grugni et al., 2012).

Based on the network analysis of haplogroup Q-M242, is that Hungarian-speaking populations living in more isolated areas (no admixture), such as Csángó and Székely males in Transylvania, Romania, clustered on the same branch as the ancient Xiongnu people lived in Inner Asia about 2000 years ago, suggesting these males are genetically related (see Figure 5). An Uzbek sample from Khwarezm (Biró et al., 2015) is also on this branch, in addition to this study's R80 sample, which shares the same haplotype as an Uzbek sample from China (Huang et al., 2018), implying a common genetic trace. Neparáczki et al., 2019 reported that, in Hungary, haplogroups Q-M25 and Q-F1096 (xM25) were found in the Hunnic period (fifth century CE) and the Hungarian Conquest period (9th-11th centuries CE), respectively. They state that although Q-M25 is rare in Europe its highest frequency is among Hungarian speaking Székely population in Transylvania, Romania. Furthermore, the authors noted that ancient samples with haplogroup Q1a2-M25 are known from the Bronze Age Okunevo and Karasuk cultures, as well as from Middle Age Tian Shan Huns and Hunnic-Sarmatians, suggesting that this lineage may be of Hunnic origin in Europe. This is confirmed by the Hun/1 sample, derived from Transylvania (Neparáczki et al., 2019). These observations also confirm the genetic relationship between contemporary Hungarians (Székely and Csángó) and the ancient Xiongnu people of Inner Asia that we found in our study.

The remaining two Rétköz Q samples (R101 and R58) cluster with Central (Uzbek, Tajik, and Turkmen) and Inner Asian (Mongolian) samples in a subbranch of the main branch, which indicates that these samples are distant from the Hungarian-Xiongnu subbranch but that they nevertheless originate from a common branch (Figure 5).

Another interesting observation is that three Hungarian males on the Hungarian-Xiongnu branch (Figure 5) form a cluster with Balkarian males in another network (see Results, data not shown). The Balkars identify as a Turkic people and speak the same language as the Karachays from Karachay-Cherkessia. Balkars and Karachay are sometimes referred to as a single ethnicity (Джантуева 2010). Our comparative phylogenetic study of folk music and genetics, indicate that the
Volga–Sicilian–Turkish–Karachay–Hungarian–Finnish–Dakota and the Chinese–Mongol–Volga–Sicilian
–Turkish–Karachay–Hungarian cultures have the largest sets of common melody types, suggesting the existence of a common "parent language" from which their music evolved. The results may also show that the populations that incorporate this musical style into their culture have common genetic roots, in particular that the development of this musical culture in their early history may be attributed to their common genetic ancestors (Pamjav et al., 2012). In this study, the similar melodies in Hungarian and Karachay folk music may suggest their common genetic traces. Unfortunately, however, comprehensive genetic studies on Balkars/Karachays are not available.

4.1.5 Median -joining network of 189 G2a-L156 haplotypes

Haplogroup G isassociated with the spread of agriculture, particularly in Europe. Haplogroup G was first discovered in Europe and Georgia (Semino et al., 2000) and was later detected in Caucasian and Hungarian populations (Nasidze et al., 2004; Völgyi et al., 2009). The frequency of haplogroup G2a-P15 (L156) is about 4% in the general Hungarian population (Völgyi et al., 2009), 6% in the Hungarian-speaking Csángó population (Transylvania, Romania), and 4% in the Hungarian-speaking Székely population (Transylvania, Romania) (Bíró et al., 2015). As the frequency of the G haplogroup is low among Hungarian-speakers, only the L156 SNP was tested from the downstream SNPs, which is phylogenetically equivalent to SNP P287 marker (www.phylotree.org/Y/tree/G.htm). Rootsi et al. analyzed 113 Hungarian males, of which 2 belonged to haplogroup G-M201 (1.8%) and to subgroups G-L497 (0.9%) and G-M406 (0.9%). These authors showed that SNP P303 defines the most frequent and widespread subhaplogroup G, whereas P303-related L497 lineages occur in Europe, where they likely originated. The highest frequency of G2a-P303 is detected in populations from the Caucasus, specifically among South Caucasian Abkhazians (24%), Northwest Caucasian Adyghe (39.7%), and Cherkessians (36.5%) (Rootsi et al., 2012). Another frequent subclade is M406, which is the sister clade of P303. The G2a-M406 has a peak frequency in the Mediterranean and Central Anatolian (6–7%) populations, as well as in Greek (4%) and Italian (3%) populations. It is not detected in many other regions with high P303 frequency (Rootsi et al., 2012). The G2a-P16 (P18) lineage is specific to the Caucasus, accounts for a third of the Caucasian male gene pool and has a high frequency in the Southern and Northwestern Caucasus, with the highest frequency among North Ossetians (63.6%). Outside the Caucasus (Anatolia, Armenia, Russia, and Spain), the P16 lineage is either present at less than 1% or is absent (Rootsi et al., 2012).

Based on our network analysis, four subclades can be distinguished as L497, P303, M406, and P18 (under P16), as the samples with known haplogroups in the network are clustered on the same branches. Hungarians are included in L497, P303, and M406 subclades, except for P18 (1 Csángo sample), indicating that the gene flow from the P18 subclade of the Caucasus has been negligible. We included nine samples from the Rétköz and Váh valley populations, each of which falls into these three subgroups. The results indicate a close genetic relationship between the Hungarian-Balkar/Karachay and Hungarian-Avar males, because they share common haplotypes (cluster 1–3) or cluster in similar haplotypes (P303, L97, and M406). Thus, these observations likely suggest a common genetic origin, rather than coicidence.

The results of the aDNA studies in the Carpathian Basin show the haplogroup G2a-L156. One ancient Avar and four Hungarian Conqueror samples belong to the subgroup G2a-L293, two Hungarian Conqueror samples belong to haplogroup G2a-U1, and one ancient Hungarian sample belongs to haplogroup G2a-L30 (Neparáczki et al., 2019; Fóthi et al., 2020). SNPs L293 and L30 are downstream of L156 but upstream of M406. U1 and L497 are sister clades of P303 under M406 SNP (http://www.phylotree.org/Y/tree/G.htm). This indicates that SNPs L293 and L30 are older than SNPs M406 and P303, whereas SNPs U1 and L497 are the youngest.

As such, the modern and ancient DNA results support the case for a common genetic origin of the Hungarian and Caucasion populations.

4.1.6 Median -joining network of 183 E1b1-M78 haplotypes

Haplogroup E, defined by mutation M40 (M96, P29), is the most common human Y-chromosome clade in Africa (www.phylotree.org/tree/E/htm). From downstream SNPs of haplogroup E, the E1b1-M35 mutation and M78 below it are two of the most frequent markers in European males. E-V13, a single clade within E, highlights a series of expansions duirng the Bronze Age in Southern Europe (Cruciani et al., 2007). All European M78 chromosomes belong to subclade V13, as well (www.phylotree.org/tree/E/htm). The highest frequency values for the M78 lineage are detected in populations from the Balkans, such as Albanians (32.29%), Bulgarians (16.67%), Macedonians (18.18%), continental Greeks (19.05%), and Southern Italians (13.07%) (Cruciani et al., 2007). The frequency of haplogroup E1b1-M78 is about 4.2% in the general Hungarian population (Völgyi et al., 2009) and 9.43% in the Hungarians analyzed by Cruciani et al. (2007). However, its presence is negligible in Caucasian populations, like the Lezghins, Ossets-Digor, Ossets-Iron, Abkhazians, Shapsugs, and Circassians (Balanovsky et al., 2011), and it is very rare in Central and Inner Asia (Biró et al., 2015).

As shown in Figure 7, the E1b1-M78 chromosomes of Hungarian-speaking populations originated in the Balkans and were introduced into Hungarian gene pool. Their location on the outer edge of the network, indicating many mutational steps, suggests that the M78 chromosomes in the Váh valley population appear to have separated from the common European ancestor earlier than those of the Rétköz population.

In aDNA studies, E1b1-M78 was observed among the Middle and Late period Avar (650–710 CE), as well as among the Hungarian Conquerors (895-mid *X*th century) (Neparáczki et al., 2019). However, the absence of STR precludes its inclusion in the network analysis.

Based on the network topologies, the most diverse haplogroups are Q-M242, G2a-L156 and R1b-L23, which can be seen from the pattern of the networks, that is, the number of unique haplotypes of a male within the network is significant as well as the age of the accumulated STR variation (TMRCA) also supports it. The divergence times of haplogroups N-M46 and R1a-Z93 are almost the same (TMRCA) and within the network there are many clusters where several males share the same haplotype. Haplogroup E-M78 is the youngest, as the number of unique haplotypes is less than in the other networks.

4.2 Genetic structure

Based on Fst analysis, the Váh valley population shows a stronger genetic similarity to Balkan populations than does the Rétköz population. This is primarily due to the relatively high E and I2a haplogroup frequencies. Both haplogroups are common in Balkan populations but are less frequently found in Eastern Europeans (Regueiro et al., 2012; Doğan et al., 2016). The Rst-based MDS plot supports this observation, as the Váh valley popular was closer to a Balkan population (Croatia), whereas the nearest population to Rétköz is a Central European population (Czech). A further difference between the two studied Hungarian populations can be concluded from the Fst-based MDS plot. It shows that the Váh valley, which is at the western part of the Hungarian-Slovakian contact zone, is genetically closer to the Western Europeans and that Rétköz is at the eastern part of that zone and is therefore genetically closer to the Eastern Europeans. This study did not detect a genetic linkage between the Váh valley and the Hungarian Conquest period populations, and the results also reveal the Váh valley's relative isolation from neighboring populations.

In summary, the results obtained by us, which show that the genetic relationships between modern Hungarians, Hungarian Conquerors, Asian Huns (Xiongnu) and ancient Avars are continuous, are fully supported by the results of the whole genome sequencing data performed by Hungarian researchers (Maróthi et al., 2022).

5 Conclusion

The genetic composition of the Rétköz (Hungary) and Váh valley (Slovakia) populations indicate different histories. In the Rétköz population, the paternal lineages that were also found in the Hungarian Conquerors, such as haplogroups R1a-Z93, N-M46, Q-M242, R1b-L23, and G2a-L156, were better preserved. The genetic composition of the Váh valley population is similar to that of the surrounding Indo-European populations.

The Hungarian males shared common haplotypes with ancient Xiongnu, ancient Avar, Caucasian Avar, Abkhazian, Balkarian, and Circassian males within haplogroups R1a-Z93, N1c-M46, and R1b-L23, suggesting a common genetic footprint. Additionally, Hungarians cluster on a common branch with the ancient Asian Huns (Xiongnu), ancient Avars, and several modern Caucasian populations (Avar, Ossetian, Balkars) within the haplogroups R1a-Z93, R1b-L23, and Q-M242, implying a close genetic relationship.

Further studies are needed to clarify if the common genetic footprints were acquired directly or indirectly. Comprehensive studies from European, Central Asian, and Caucasian populations should be conducted for the haplogroups using more downstream SNPs and NGS sequencing to learn more about the origins, expansion, and ethno-linguistic affiliations of the populations.

Data availability statement

The data presented in the study can be found in the YHRD respository, accession numbers: YA004754 and YA0047555.

Ethics statement

The studies involving human participants were reviewed and approved by the according to the recommenndations of the Data Protection Code the Humanities Research Center (MTA BTK-KP/450-17/2018). The patients/participants provided their written informed consent to participate in this study.

Author contributions

HP and EF designed the study and conducted the experiment. HP wrote the manuscript. EF and AF collected the samples. AT, DD, EF, VK, and HP analyzed the data. AT, DD and AF visualized data. All authors reviewed the manuscript. HP and EF contributed equally to this work.

Funding

This work was supported by the Thematic Excellence Program (TKP20-NKA-020) of the Ministry of Innovation and Technology in Hungary.

Acknowledgments

We are grateful to Tamás Oláh and György Mészáros for their valuable help in collecting the samples, as well as to Ariana Gugora for proofreading the manuscript in her native English. We also thank all the sample donors and the laboratory assistants.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.977517/full#supplementary-material

References

Adamov, D., Guryanov, V., Karzhavin, S., Tagankin, V., and Urasin, V. (2015). Defining a new rate constant for Y-chromosome SNPs based on full sequencing data. *Russ. J. Genet. Genealogy (*Русская версия*):*Том 7, 1920–2997.

Balanovsky, O., Dibirova, K., Dybo, A., Mudrak, O., Frolova, S., Pocheshkhova, E., et al. (2011). Parallel evolution of genes and languages in the Caucasus region. *Mol. Biol. Evol.* 28 (10), 2905–2920. doi:10.1093/molbev/msr126

Balloux, F., and Goudet, J. (2002). Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Mol. Ecol.* 11 (4), 771–783. doi:10.1046/j.1365-294x.2002.01474.x

Bandelt, H. J., Forster, P., Röhl, A., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi:10.1093/oxfordjournals.molbev.a026036

Batini, C., Hallast, P., Vågene, Å. J., Zadik, D., Eriksen, H. A., Pamjav, H., et al. (2017). Population resequencing of European mitochondrial genomes highlights sex-bias in Bronze Age demographic expansions. *Sci. Rep.* 7 (1), 12086. doi:10.1038/s41598-017-11307-9

Battaglia, V., Fornarino, S., Al-Zahery, N., Olivieri, A., Pala, M., Myres, N. M., et al. (2009). Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. *Eur. J. Hum. Genet.* 17 (6), 820–830. doi:10.1038/ejhg.2008.249

Bíró, A., Fehér, T., Bárány, G., and Pamjav, H. (2015). Testing central and inner asian admixture among contemporary Hungarians. *Forensic Sci. Int. Genet.* 15, 121–126. doi:10.1016/j.fsigen.2014.11.007

Bogácsi-Szabó, E., Kalmár, T., Csányi, B., Tömöry, G., Czibula, A., Priskin, K., et al. (2005). Mitochondrial DNA of ancient cumanians: Culturally asian steppe nomadic immigrants with substantially more Western eurasian mitochondrial DNA lineages. *Hum. Biol.* 77 (5), 639–662. doi:10.1353/hub.2006.0007

Brandstätter, A., Egyed, B., Zimmermann, B., Duftner, N., Padar, Z., Parson, W., et al. (2007). Migration rates and genetic structure of two Hungarian ethnic groups in Transylvania, Romania. *Ann. Hum. Genet.* 71, 791–803. doi:10.1111/j.1469-1809.2007.00371.x

Cruciani, F., La, F. R., Trombetta, B., Santolamazza, P., Sellitto, D., Colomb, E. B., et al. (2007). Tracing past human male movements in northern/eastern Africa and Western Eurasia: New clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol. Biol. Evol.* 24 (6), 1300–1311. doi:10.1093/molbev/msm049

Csáky, V., Gerber, D., Koncz, I., Csiky, G., Mende, B. G., Szeifert, B., et al. (2020). Author Correction: Genetic insights into the social organisation of the Avar period elite in the 7th century AD Carpathian Basin. *Sci. Rep.* 10 (1), 13398. doi:10.1038/s41598-020-69583-x

Derenko, M. V., Maliarchuk, B. A., Wozniak, M., Denisova, G. A., Dambueva, I. K., Dorzhu, C. M., et al. (2007). Distribution of the male lineages of Genghis Khan's descendants in northern Eurasian populations. *Russ. J. Genet.* 43 (3), 334–337. doi:10.1134/s1022795407030179

Di, C. J., Pennarun, E., Mazieres, S., Myres, N. M., Lin, A. A., Temori, S. A., et al. (2013). Afghan hindu kush: Where eurasian sub-continent gene flows converge. *PLoS One* 8 (10), e76748. doi:10.1371/journal.pone.0076748

Doğan, S., Ašić, A., Doğan, G., Besic, L., and Marjanovic, D. (2016). Y-chromosome haplogroups in the Bosnian-Herzegovinian population based on 23 Y-STR Loci. *Hum. Biol.* 88 (3), 201–209. doi:10.13110/humanbiology.88.3.0201

Dudás, E., Váhó-Zalán, A., Vándor, A., Saypasheva, A., Pomozi, P., and Pamjav, H. (2019). Genetic history of bashkirian Mari and southern Mansi ethnic groups in the Ural region. *Mol. Genet. Genomics* 294 (4), 919–930. doi:10.1007/s00438-019-01555-x

Ehaotufca, V. R. (2010). Urpxfss j ;taV9 vprnjrpcaoj>larayafcp-bamlarslp[p ;topsa j raicjtjf ;tojyfslpk j rfmj[jpiopk jefotjyopstj. dissercat.

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under linux and windows. *Mol. Ecol. Resour.* 10 (3), 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Fehér, T., Németh, E., Vándor, A., Kornienko, I. V., Csáji, L. K., Pamjav, H., et al. (2015). Y-SNP L1034: Limited genetic link between Mansi and Hungarian-speaking populations. *Mol. Genet. Genomics* 290 (1), 377–386. doi:10.1007/s00438-014-0925-2

Fóthi, E., Gonzalez, A., Fehér, T., Gugora, A., Fóthi, A., Biró, O., et al. (2020). Genetic analysis of male Hungarian conquerors: European and asian paternal lineages of the conquering Hungarian tribes. *Archaeol. Anthropol. Sci.* 12, 31. doi:10.1007/s12520-019-00996-0

Gayden, T., Cadenas, A. M., Regueiro, M., Singh, N. B., Zhivotovsky, L. A., Underhill, P. A., et al. (2007). The himalayas as a directional barrier to gene flow. *Am. J. Hum. Genet.* 80 (5), 884–894. doi:10.1086/516757

Grugni, V., Battaglia, V., Hooshiar, K. B., Parolo, S., Al-Zahery, N., Achilli, A., et al. (2012). Ancient migratory events in the Middle East: New Clues from the Y-chromosome variation of modern Iranians. *PLoS One* 7 (7), e41252. doi:10.1371/journal.pone.0041252

Huang, Y. Z., Pamjav, H., Flegontov, P., Stenzl, V., Wen, S. Q., Tong, X. Z., et al. (2018). Dispersals of the siberian Y-chromosome haplogroup Q in Eurasia. *Mol. Genet. Genomics* 293 (1), 107–117. doi:10.1007/s00438-017-1363-8

Ilumäe, A. M., Reidla, M., Chukhryaeva, M., Järve, M., Post, H., Karmin, M., et al. (2016). Human Y chromosome haplogroup N: A non-trivial time-resolved phylogeography that cuts across language families. *Am. J. Hum. Genet.* 99 (1), 163–173. doi:10.1016/j.ajhg.2016.05.025

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L., and Hammer, M. F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18, 830–838. doi:10.1101/gr.7172008

Keyser, C., Zvénigorosky, V., Gonzalez, A., Fausser, J. L., Jagorel, F., Gérard, P., et al. (2021). Genetic evidence suggests a sense of family, parity and conquest in the Xiongnu Iron Age nomads of Mongolia. *Hum. Genet.* 140 (2), 349–359. doi:10.1007/s00439-020-02209-4

Liptak, P. (1979). *Magyar etnogenezis*. Budapest, Tankönyvkiadó: Embertan és emberszármazástan, 317–319.

Maár, K., Varga, G. I. B., Kovács, B., Schütz, O., Maróti, Z., Kalmár, T., et al. (2021). Maternal lineages from 10-11th century commoner cemeteries of the Carpathian Basin. *Genes. (Basel)* 12 (3), 460. doi:10.3390/genes12030460

Malyarchuk, B., Derenko, M., Denisova, G., Maksimov, A., Wozniak, M., Grzybowski, T., et al. (2011). Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a. *J. Hum. Genet.* 56 (8), 583–588. doi:10.1038/jhg.2011.64

Maróti, Z., Neparáczki, E., Schütz, O., Maár, K., Varga, G. I. B., Kovács, B., et al. (2022). The genetic origin of Huns, Avars, and conquering Hungarians. *Curr. Biol.* 32, 2858–2870.e7. doi:10.1016/j.cub.2022.04.093

Menges, K. H. (2011). *Altaic.* Encyclopædia Iranica.

Myres, N. M., Rootsi, S., Lin, A. A., Järve, M., King, R. J., Kutuev, I., et al. (2011). A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* 19 (1), 95–101. doi:10.1038/ejhg.2010.146

Nagy, P. L., Olasz, J., Neparáczki, E., Rouse, N., Kapuria, K., Cano, S., et al. (2021). Determination of the phylogenetic origins of the Árpád Dynasty based on Y chromosome sequencing of Béla the Third. *Eur. J. Hum. Genet.* 29 (1), 164–172. doi:10.1038/s41431-020-0683-z

Nasidze, I., Ling, E. Y., Quinque, D., Dupanloup, I., Cordaux, R., Rychkov, S., et al. (2004). Mitochondrial DNA and Y-chromosome variation in the caucasus. *Ann. Hum. Genet.* 68, 205–221. doi:10.1046/j.1529-8817.2004.00092.x

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* 70 (12), 3321–3323. doi:10.1073/pnas.70.12.3321

Neparáczki, E., Juhász, Z., Pamjav, H., Fehér, T., Csányi, B., Zink, A., et al. (2017a). Genetic structure of the early Hungarian conquerors inferred from mtDNA haplotypes and Y-chromosome haplogroups in a small cemetery. *Mol. Genet. Genomics* 292 (1), 201–214. doi:10.1007/s00438-016-1267-z

Neparáczki, E., Kocsy, K., Tóth, G. E., Maróti, Z., Kalmár, T., Bihari, P., et al. (2017b). Revising mtDNA haplotypes of the ancient Hungarian conquerors with next generation sequencing. *PLoS One* 12 (4), e0174886. doi:10.1371/journal.pone.0174886

Neparáczki, E., Maróti, Z., Kalmár, T., Kocsy, K., Maár, K., Bihari, P., et al. (2018). Mitogenomic data indicate admixture components of Central-Inner Asian and Srubnaya origin in the conquering Hungarians. *PLoS One* 13 (10), e0205920. doi:10.1371/journal.pone.0205920

Neparáczki, E., Maróti, Z., Kalmár, T., Maár, K., Nagy, I., Latinovics, D., et al. (2019). Y-chromosome haplogroups from Hun, Avar and conquering Hungarian period nomadic people of the Carpathian Basin. *Sci. Rep.* 9 (1), 16569. doi:10.1038/s41598-019-53105-5

Olasz, J., Seidenberg, V., Hummel, S., Szentirmay, Z., Szabados, G., Melegh, B., et al. (2018). DNA profiling of Hungarian King Béla III and other skeletal remains originating from the Royal Basilica of Székesfehérvár. *Archaeol. Anthropol. Sci.* 11, 1345–1357. doi:10.1007/s12520-018-0609-7

Pamjav, H., Fóthi, Á., Fehér, T., and Fóthi, E. (2017). A study of the Bodrogköz population in north-eastern Hungary by Y chromosomal haplotypes and haplogroups. *Mol. Genet. Genomics* 292 (4), 883–894. doi:10.1007/s00438-017-1319-z

Pamjav, H., Juhász, Z., Zalán, A., Németh, E., and Damdin, B. (2012). A comparative phylogenetic study of genetics and folk music. *Mol. Genet. Genomics* 287 (4), 337–349. doi:10.1007/s00438-012-0683-y

Pimenoff, V. N., Comas, D., Palo, J. U., Vershubsky, G., Kozlov, A., and Sajantila, A. (2008). Northwest siberian khanty and Mansi in the junction of west and east eurasian gene pools as revealed by uniparental markers. *Eur. J. Hum. Genet.* 16 (10), 1254–1264. doi:10.1038/ejhg.2008.101

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Regueiro, M., Rivera, L., Damnjanovic, T., Lukovic, L., Milasin, J., and Herrera, R. J. (2012). High levels of Paleolithic Y-chromosome lineages characterize Serbia. *Gene* 498 (1), 59–67. doi:10.1016/j.gene.2012.01.030

Róna-Tas, A. (1999). *Hungarians and Europe in the early Middle ages: An introduction to early Hungarian history*. Budapest, Hungary: Central European University Press.

Rootsi, S., Myres, N. M., Lin, A. A., Järve, M., King, R. J., Kutuev, I., et al. (2012). Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the caucasus. *Eur. J. Hum. Genet.* 20 (12), 1275–1282. doi:10.1038/ejhg.2012.86

Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., et al. (2000). The genetic legacy of paleolithic *Homo sapiens* sapiens in extant Europeans: A Y chromosome perspective. *Science* 290, 1155–1159. doi:10.1126/science.290.5494.1155

Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C-E. T., et al. (2006). Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.* 78, 202–221. doi:10.1086/499411

Tömöry, G., Csányi, B., Bogácsi-Szabó, E., Kalmár, T., Czibula, A., Csosz, A., et al. (2007). Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. *Am. J. Phys. Anthropol.* 134 (3), 354–368. doi:10.1002/ajpa.20677

Underhill, P. A., Poznik, G. D., Rootsi, S., Järve, M., Lin, A. A., Wang, J., et al. (2015). The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *Eur. J. Hum. Genet.* 23 (1), 124–131. doi:10.1038/ejhg.2014.50

Völgyi, A., Zalán, A., Szvetnik, E., and Pamjav, H. (2009). Hungarian population data for 11 Y-STR and 49 Y-SNP markers. *Forensic Sci. Int. Genet.* 3 (2), e27–e28. doi:10.1016/j.fsigen.2008.04.006

Willuweit, S., and Roewer, L. (2015). The new Y Chromosome haplotype reference database. *Forensic Sci. Int. Genet.* 15, 43–48. doi:10.1016/j.fsigen.2014.11.024

Zhao, Y-B., Zhang, Y., Zhang, Q-C., Li, H-J., Cui, Y-Q., Xu, Z., et al. (2015). Ancient DNA reveals that the genetic structure of the northern han Chinese was shaped prior to 3, 000 years ago. *PLoS One* 10 (5), e0125676. doi:10.1371/journal.pone.0125676

Zhivotovsky, L. A., Underhill, P. A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T., et al. (2004). The effective mutation rate at Y chromosome Short Tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* 74 (1), 50–61. doi:10.1086/380911

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fgene.2022.1030463

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Chenqi Wang,
University of South Florida,
United States
Edith Christiane Bougouma,
Groupe de Recherche Action en Santé
(GRAS), Burkina Faso

*CORRESPONDENCE
Kristan Alexander Schneider ,
kristan.schneider@hs-mittweida.de

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

RECEIVED 29 August 2022
ACCEPTED 26 September 2022
PUBLISHED 03 November 2022

CITATION
Schneider KA and Salas CJ (2022),
Evolutionary genetics of malaria.
Front. Genet. 13:1030463.
doi: 10.3389/fgene.2022.1030463

Evolutionary genetics of malaria

Kristan Alexander Schneider[1]* and Carola Janette Salas[2]

[1]Department of Applied Computer- and Biosciences, University of Applied Sciences Mittweida, Mittweida, Germany, [2]Department of Parasitology, U.S. Naval Medical Research Unit No 6 (NAMRU-6), Lima, Peru

Many standard-textbook population-genetic results apply to a wide range of species. Sometimes, however, population-genetic models and principles need to be tailored to a particular species. This is particularly true for malaria, which next to tuberculosis and HIV/AIDS ranks among the economically most relevant infectious diseases. Importantly, malaria is not one disease—five human-pathogenic species of *Plasmodium* exist. *P. falciparum* is not only the most severe form of human malaria, but it also causes the majority of infections. The second most relevant species, *P. vivax*, is already considered a neglected disease in several endemic areas. All human-pathogenic species have distinct characteristics that are not only crucial for control and eradication efforts, but also for the population-genetics of the disease. This is particularly true in the context of selection. Namely, fitness is determined by so-called fitness components, which are determined by the parasites live-history, which differs between malaria species. The presence of hypnozoites, i.e., dormant liver-stage parasites, which can cause disease relapses, is a distinct feature of *P. vivax* and *P. ovale* sp. In *P. malariae* inactivated blood-stage parasites can cause a recrudescence years after the infection was clinically cured. To properly describe population-genetic processes, such as the spread of anti-malarial drug resistance, these features must be accounted for appropriately. Here, we introduce and extend a population-genetic framework for the evolutionary dynamics of malaria, which applies to all human-pathogenic malaria species. The model focuses on, but is not limited to, the spread of drug resistance. The framework elucidates how the presence of dormant liver stage or inactivated blood stage parasites that act like seed banks delay evolutionary processes. It is shown that, contrary to standard population-genetic theory, the process of selection and recombination cannot be decoupled in malaria. Furthermore, we discuss the connection between haplotype frequencies, haplotype prevalence, transmission dynamics, and relapses or recrudescence in malaria.

KEYWORDS

complexity of infection (COI), co-infection, mixed-species infection, recrudescence, relapse, seed bank, hypnozoites, multiplicity of infection (MOI)

1 Introduction

After a decade of declining incidence the number of annual malaria infections rises since 2018, challenging the WHO goal to reduce malaria incidence by at least 90% by 2030 (WHO, 2021a). This is partly attributed to the rapid emergence and spread of anti-malarial drug resistance, an evolutionary-genetic process whose understanding is a global health priority (WHO, 2021b).

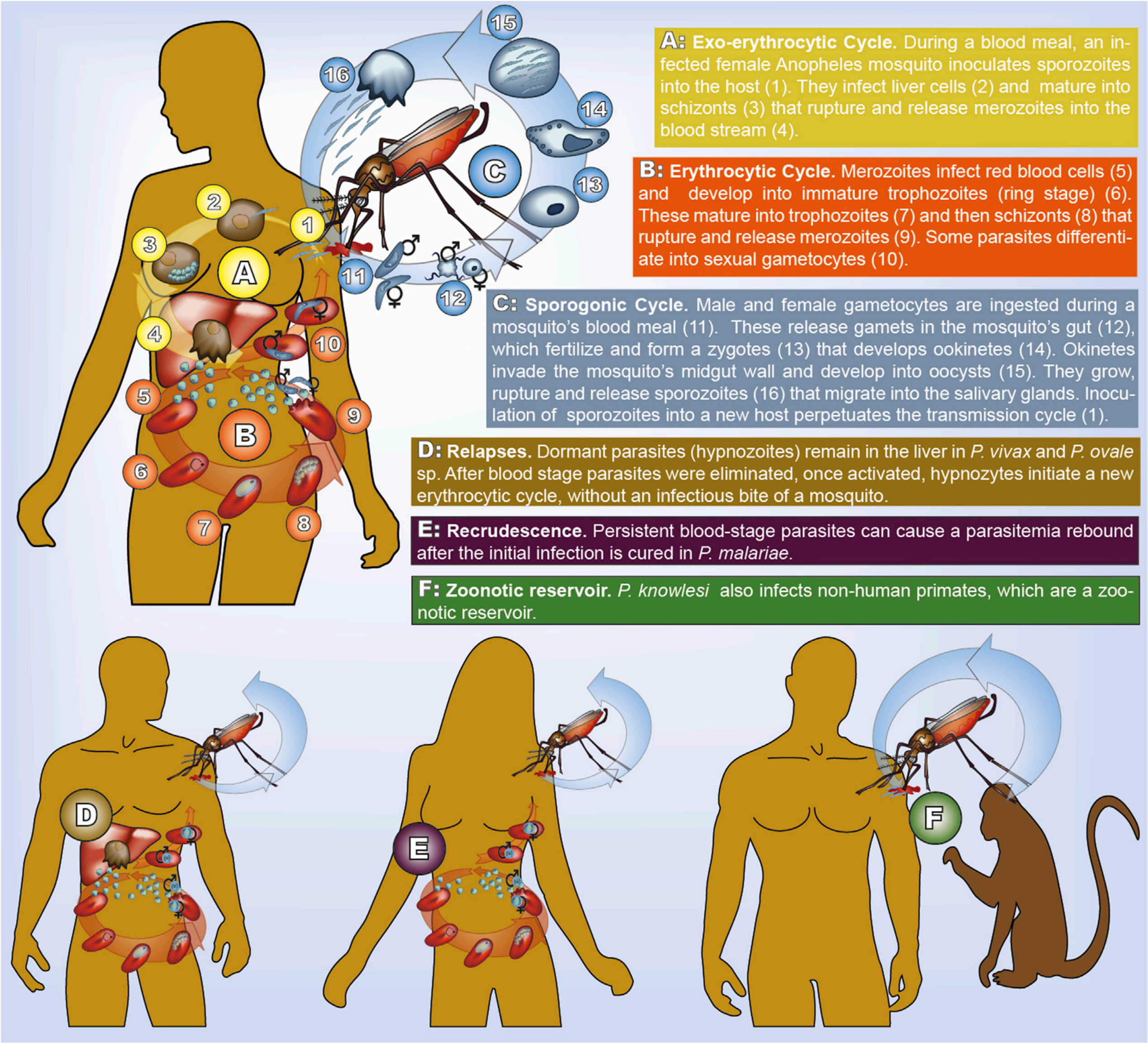
Malaria is caused in humans and animals by *Plasmodium* parasites. These unicellular, haploid eukaryotes are transmitted by numerous species of female *Anopheles* mosquitoes. Both the parasite and vector species are adapted to specific human or animal hosts. Five species of *Plasmodium* are pathogenic to humans, which can be transmitted by over 100 *Anopheles* species (Nicoletti, 2020). Over 95% of the 240 million annual infections and 620,000 deaths worldwide are attributed to *P. falciparum*. Although, the WHO recommended the use of RTS,S, the first approved malaria vaccine, in children to prevent *P. falciparum* infections in areas of moderate to high transmission, the vaccine's efficacy is low and malaria control depends strongly on reliable diagnostics and drug treatments to cure acute infections (Greenwood et al., 2021). While the second most relevant species, *Plasmodium vivax*, receives considerable attention, the other species *P. ovale* sp., *P. malariae*, and *P. knowlesi* are somewhat neglected, due to an outdated distinction between harmful and harmless malaria species (Lover et al., 2018).

The spread of deletions in the histidine-rich protein 2 and 3 (HRP2/3) genes of *P. falciparum*, which encode for the antigens targeted by rapid diagnostic tests (RDTs) as well as drug-resistant *P. falciparum* and *P. vivax* haplotypes substantially challenge successful malaria control. These evolutionary genetic processes are tightly linked to the pathogen's complex transmission cycle, which besides some species-specific differences, is commonly shared among all *Plasmodia* (Su et al., 2019; Beshir et al., 2022).

The transmission cycle starts with an infected mosquito taking her blood meal. She inoculates parasites in the form of sporozoites from her salivary glands into the human body. This is followed by the exo-erythrocytic cycle, during which sporozoites reach the liver to infect hepatocytes. In the infected liver cells parasites mature into schizonts. The erythrocytic cycle is initiated when the schizonts rupture and merozoites are released into the bloodstream. Erythrocytes are invaded by merozoites, which form ring stage trophozoites and then mature into schizonts. Once they rupture, new merozoites are released into the bloodstream. During this step of asexual reproduction, some parasites differentiate into male or female gametocytes, which do not reproduce in the human host. Once a mosquito ingests male and female gametocytes, the sporogonic cycle is initiated. Gametes released by male and female gametocytes fertilize and form zygotes. Following a step of meiosis, and hence recombination, the zygote becomes tetraploid and develops into ookinetes, which migrate through the midgut wall and transform into oocysts. In the oocyst sporozoite budding occurs in the haploid state. Division of each oocyst produces thousands of sporozoites that move into the mosquito salivary glands, completing the transmission cycle. Because gametocytes immediately release gametes, only parasites exiting the same host recombine, potentially leading to a high degree of inbreeding during the sexual reproduction of the parasite (Ngwa et al., 2016).

Species-specific differences occur in the number of parasites within an infection (parasitemia and gametocytemia counts), and the duration of the various phases in the transmission cycle. The replication of merozoites in 72-hour- rather than 48-hour-cycles distinguishes *P. ovale* sp. from other species. The onset of gametocytogenesis and the longevity of gametocytes were argued to accelerate drug-resistance evolution in *P. falciparum* compared to *P. vivax* (Schneider and Escalante, 2013). Dormant liver-stage parasites (hypnozoites), can result in disease relapses weeks, months, or even years after the clearance of blood stage parasites and occur only in *P. vivax* and *P. ovale* sp. Currently primaquine (PQ) and tafenoquine (TQ) are the only approved drugs to clear hypnozoites (Watson et al., 2021). Unfortunately, patients with (glucose-6 phosphate dehydrogenase) G6PD deficiency, which is widespread in many malaria-endemic areas, cannot be treated with these drugs (Baird et al., 2018; Dean et al., 2020). Extremely prolonged carriage of blood-stage parasites causing recrudescences occur in *P. malariae* (Collins and Jeffery, 2007). It is commonly accepted, although not completely ruled out, that the rebounce of parasitaemia in *P. malariae* is not caused by quiescent pre-erythrocytic stages such as hypnozoites. Because of relapses occurring in *P. vivax*, *P. ovale* sp., and prolonged blood stage parasite carriage in *P. malariae*, these species are resilient in areas in which *P. falciparum* transmission cannot be sustained. While all other human malaria species can—at least in theory—be eradicated by concentrating on the human host, this is not possible for *P. knowlesi*, which is characterized by zoonotic transmission. It became the predominant species in several endemic countries in Southeast Asia, which shifted from malaria control toward elimination (Sutherland, 2016).

The characteristics of the transmission cycle render the application of standard textbook population-genetic results incorrect. Particularly it was shown that the process of selection acting on parasites in the human hosts (including selection for drug resistance) and recombination cannot be separated (Schneider and Kim, 2010). Hence, population-genetic theory and models have to be tailored to the malaria transmission cycle. This has been done mainly for *P. falciparum*. Because a clear path to eradication has been chartered only for *P. falciparum*, the other malaria species gain more importance due to their resilient nature (Lover et al., 2018). This requires to further adapt population-genetic theory to the characteristics of other human-pathogenic malaria species.



A: Exo-erythrocytic Cycle. During a blood meal, an infected female Anopheles mosquito inoculates sporozoites into the host (1). They infect liver cells (2) and mature into schizonts (3) that rupture and release merozoites into the blood stream (4).

B: Erythrocytic Cycle. Merozoites infect red blood cells (5) and develop into immature trophozoites (ring stage) (6). These mature into trophozoites (7) and then schizonts (8) that rupture and release merozoites (9). Some parasites differentiate into sexual gametocytes (10).

C: Sporogonic Cycle. Male and female gametocytes are ingested during a mosquito's blood meal (11). These release gamets in the mosquito's gut (12), which fertilize and form a zygotes (13) that develops ookinetes (14). Okinetes invade the mosquito's midgut wall and develop into oocysts (15). They grow, rupture and release sporozoites (16) that migrate into the salivary glands. Inoculation of sporozoites into a new host perpetuates the transmission cycle (1).

D: Relapses. Dormant parasites (hypnozoites) remain in the liver in *P. vivax* and *P. ovale* sp. After blood stage parasites were eliminated, once activated, hypnozytes initiate a new erythrocytic cycle, without an infectious bite of a mosquito.

E: Recrudescence. Persistent blood-stage parasites can cause a parasitemia rebound after the initial infection is cured in *P. malariae*.

F: Zoonotic reservoir. *P. knowlesi* also infects non-human primates, which are a zoonotic reservoir.

FIGURE 1
Transmission cycle of human malaria. All species have the same cycle, but parasites life-stages have different morphology (illustrated here for *P. falciparum*). In *P. vivax* and *P. ovale* sp. dormant hypnozoites remain in the liver. In *P. malariae* recrudescence form prolonged blood stage parasites occur. In *P. knowlesi* humans and non-human primates can be infected.

Here, we extend a population-genetic framework, originally developed for *P. falciparum*, to be applicable to all other malaria species.

We exemplify the importance of species-specific differences by clarifying the role of hypnozoites in the evolution of drug resistance in *P. vivax* vs. *P. falciparum*. We also clarify, how haplotype frequencies (i.e., their relative abundance in the parasite population) and prevalence (i.e., the likelihood that a given haplotype occurs in an infection) are affected by relapses/recrudescence in other malaria species. Based on this framework, we discuss past and current developments with relevance for the evolutionary genetics of malaria.

2 Methods

We extend the population-genetic framework of (Schneider and Kim, 2010; Schneider and Kim, 2011; Schneider, 2021) that describes the temporal change in the distribution of parasite haplotypes due to recombination and selection in generations of

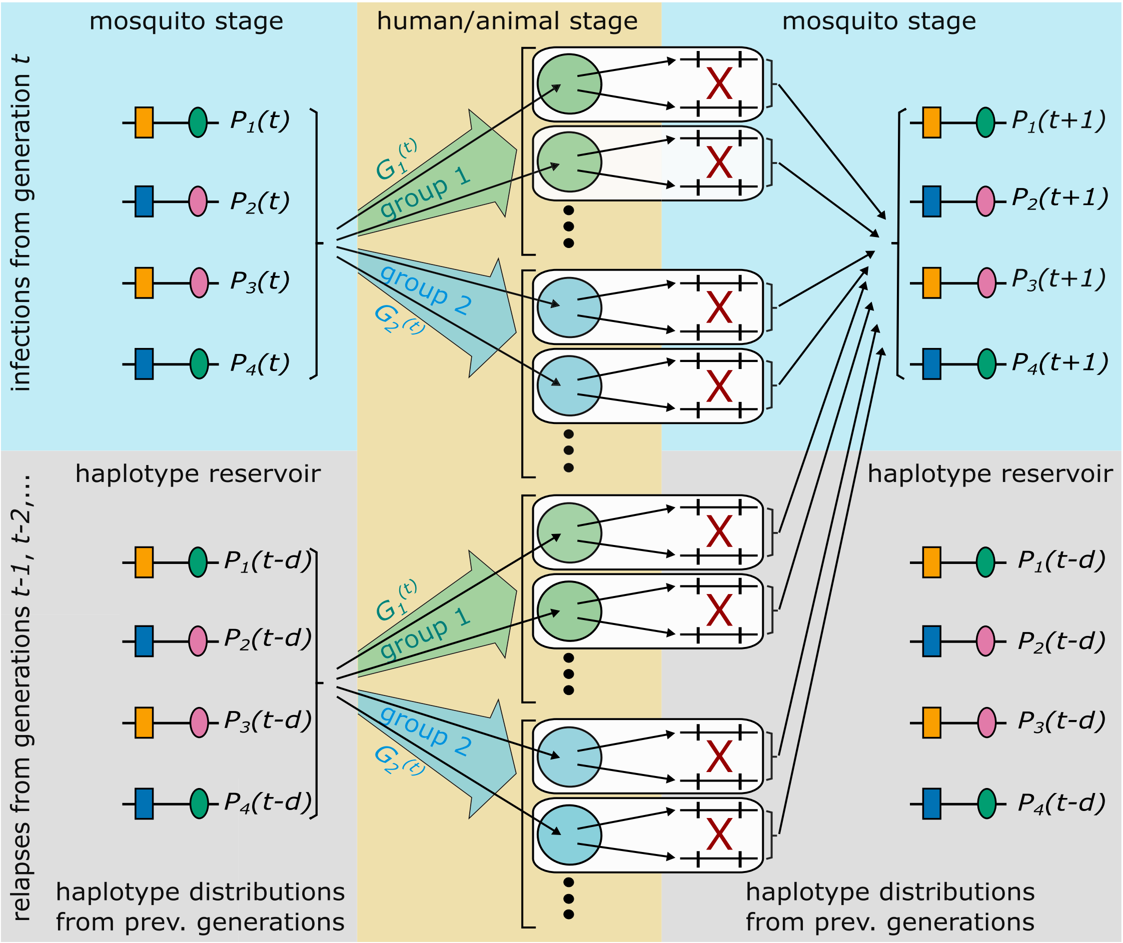


FIGURE 2
Illustrated is the idealization of the malaria transmission cycle underlying the population-genetic framework. The illustrated genetic architecture of malaria haplotypes assumes two biallelic loci, leading to four possible haplotypes. Furthermore, two groups of hosts are illustrated. Each host is infected by randomly drawing haplotypes from generation t, or a relapse/recrudescence from a previous generation occurs, which corresponds to randomly draw parasites from a previous generation (haplotype reservoir). With probability $G_g^{(t)}$ a host belongs to group g in generation t. The selective environment is different in the two groups. Recombination occurs exclusively between haplotypes exiting the same host. After recombination, haplotypes in the mosquitoes are pooled together to derive their distribution in generation t +1.

transmission cycles. While the original framework was tailored to *P. falciparum*, the extension captures the characteristics of all human-pathogenic malaria species.

The model is based on an idealization of the complex malaria transmission cycle (*cf.* Figure 1), which is illustrated in Figure 2. Although, pathogen, mosquito vector, human hosts (and, in the case of *P. knowlesi* the animal host) are involved in transmission, the framework does not require to model transmission dynamics (i.e., the interaction of mosquito vectors and human or animal hosts) explicitly. This conceptional advantages arise, because haplotype frequencies are considered at the end of the sporogenic cycle (*cf.* Figure 2). Thus, the frequency distribution of parasite haplotypes in the mosquitoes' salivary glands, which are ready for vector-host transmission, is followed.

Host and vector populations are assumed to be sufficiently large and malaria infections sufficiently frequent to justify a deterministic description of the evolutionary dynamics. Steps of full transmission cycles correspond to steps of sexual reproduction, because only one step of sexual reproduction occurs during one full transmission cycle, namely inside the mosquito vector. Many steps of asexual reproduction occur inside the vectors and hosts.

2.1 Genetic architecture of haplotypes

The genetic architecture of haplotypes is determined by their allelic configuration at one or several loci. We denote the number of all possible haplotypes by H. E.g., L biallelic loci lead to $H = 2^L$ haplotypes. In general, if haplotypes are determined by L loci, and n_l alleles are segregating at locus l, $H = n_1 \cdot n_2 \cdot \ldots \cdot n_L$. The frequency of haplotype h in generation t is denoted by $P_h^{(t)}$. Collectively, the vector of haplotype frequencies is $\boldsymbol{P}_t = (P_1^{(t)}, \ldots, P_H^{(t)})$.

2.2 Idealizing the transmission cycle

The idealized transmission cycles allows to describe the evolutionary genetics of malaria in generations of full transmission cycles (Figure 2). In generation t, it is assumed that all hosts are infected (or have a relapse or recrudescence) at the same time. Moreover, host-vector transmission is also synchronized. Inside the mosquito, parasites, which were ingested by the mosquitoes, can recombine during one step of sexual reproduction. This determines the distribution of haplotypes in the mosquitoes' salivary glands of the parasite (sporozoite) population in generation $t + 1$.

2.2.1 Heterogeneity

Disease exposure and transmission intensities are heterogeneous in endemic areas and change over time (e.g. in the context of seasonal transmission) (Bousema et al., 2011; Selvaraj et al., 2018). Moreover, hosts are heterogeneous regarding their level of genetic and naturally acquired immunity, number of co-morbidities, or the drug treatment they receive to cure the infection (in case they receive any), *etc.* (Hedrick, 2011; Gonzales et al., 2020). All of these factors can be addressed by modeling hosts in different groups (strata). Let $G_g^{(t)}$ be the probability that a host, in which an infection occurs in generation t, belongs to group g. Hence, $G_1^{(t)} + \cdots + G_S^{(t)} = 1$ for every generation t.

The number of groups, S, has to be chosen to capture the features important to the specific application of the framework. For instance, when considering drug resistance evolution, a simple distinction would be between treated and untreated infections, i.e., $S = 2$. In the case of *P. knowlesi* different groups can model human and animal hosts. In the simplest case one would have just two groups ($S = 2$), namely humans and animals.

2.2.2 Relapses and recrudescence

Hosts are not modelled explicitly. This becomes relevant when considering relapses (in *P. vivax* and *P. ovale* sp.) and recrudescence in *P. malariae*. In the following we use relapse and recrudescence synonymously, unless a distinction is necessary.

In the idealized transmission cycle, a relapse in generation t, which occurs after a delay of d generations, is equivalent to a new infection from the sporozoite population from d generations in the past, i.e., from generation $t - d$. Let $R_d^{(t)}$ be the probability that an infection in generation t is a relapse, with a delay of d generations, where $R_0^{(t)}$ is the probability of a new infection at time t. Assuming the maximum possible delay is D, the relation $\sum_{d=0}^{D} R_d^{(t)} = 1$ for all t, and $1 - R_0^{(t)}$ is the probability that a relapse occurs at time t.

The framework models the haplotype distribution in generations of transmission cycles not in real-time. The higher the transmission intensities, the more transmission cycles occur per year. The choice of the distribution of relapses has to take this into account (see Results section The effect of recrudescences and relapses). Moreover, the timing of relapses depends on the *Plasmodium* species (White, 2011).

Importantly, a host might have been exposed differently to the disease in the past, i.e., the host might belong to different groups in generations $t - d$ and t. Let $G_{g',g}^{(t-d,t)}$ be the probability that a host, who belonged to group g' in generation $t - d$, belongs to group g in generation t ($d \geq 0$). Marginalisation yields

$$G_g^{(t)} = \sum_{g'=1}^{S} G_{g',g}^{(t-d,t)} \tag{1}$$

for all t, d, g. Hence, the probability that a relapse occurs in generation t in a host in group g after a delay of d generations, when he belonged to group g', is given by

$$R_d^{(t)} G_{g',g}^{(t-d,t)}.$$

2.3 Vector-host transmission and multiplicity of infection

The presence of multiple genetically distinct parasite haplotypes within an infection is frequently referred to as multiplicity of infection (MOI) or complexity of infections (COI) and considered important in malaria. The terms MOI and COI are ambiguously defined in the literature (see (Schneider et al., 2022) for a comprehensive review). Although, it is unclear whether MOI is affecting the clinical pathogenesis of malaria, or whether different parasite haplotypes are competing within infections (intra-host competition), MOI mediates the amount of meiotic recombination and scales with transmission intensities (Pacheco et al., 2020) (see Figure 3).

Different parasite haplotypes can occur within an infection, because they are 1) sequentially transmitted (during the course of one disease episode) by different mosquitoes (super-infection); 2) co-transmitted by one mosquito (co-infection); 3) mixed up with parasites from previous infections by relapses or recrudescence.

Concerning models of MOI, the focus was mainly on super-infections. More recently, the importance of co-infections is being emphasized. Namely, more parasite genomics data is being generated, which has enough resolution to study genetic relatedness of parasites. Such data is appropriate for molecular surveillance of transmission routes (Ndiaye et al., 2021). Formal population-genetic frameworks to describe the evolutionary

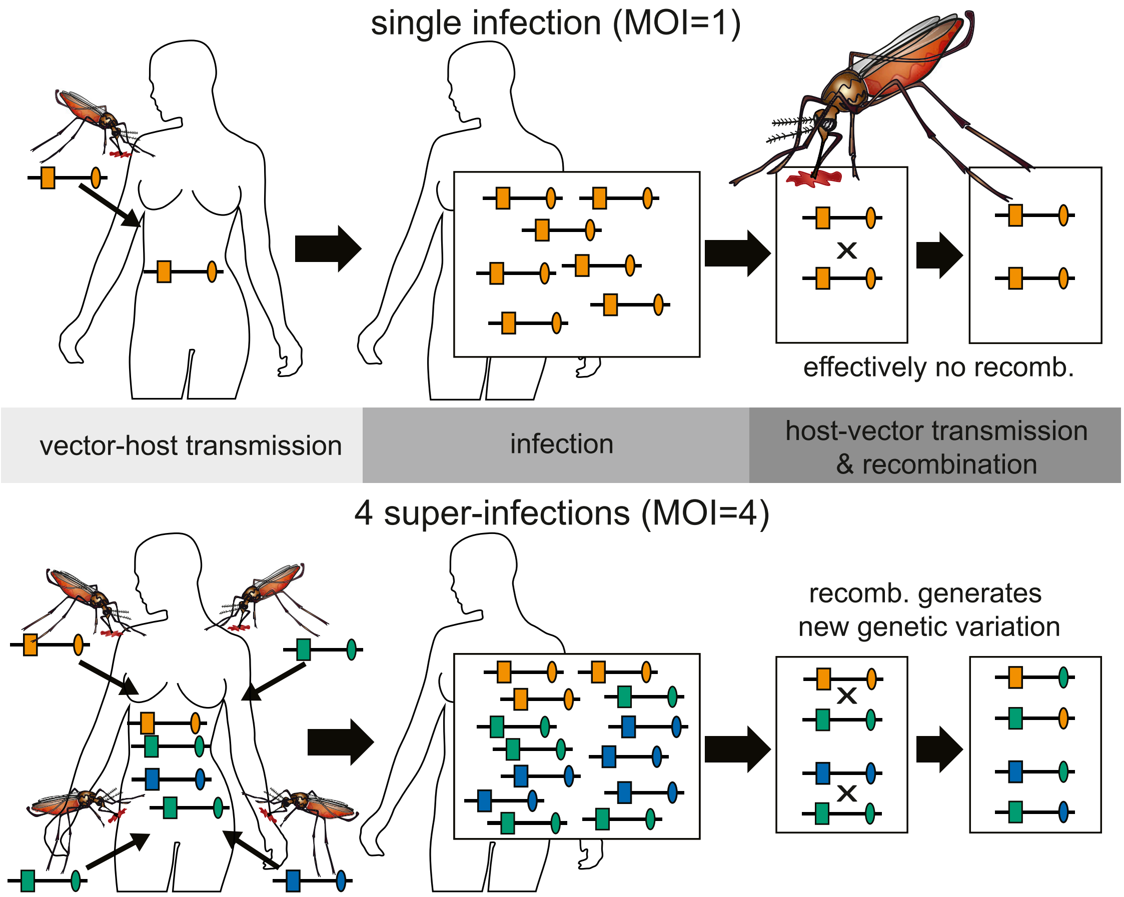

single infection (MOI=1)
effectively no recomb.
vector-host transmission
infection
host-vector transmission & recombination
4 super-infections (MOI=4)
recomb. generates new genetic variation

FIGURE 3
Illustration of the relationship between inbreeding and MOI. Top: An infection with MOI = 1 (single-clone infection) leads only to recombination between clones, i.e., effectively to no recombination. Bottom: Shown is a super-infection with four infective events (MOI = 4) and three different haplotypes being transmitted (one haplotype is transmitted independently by two mosquitoes). Recombination between the illustrated haplotypes leads to the creation of new haplotypes.

genetics of malaria that consider relapses do not exist. Mathematical models describing relapses in *P. vivax* and *P. ovale* sp. are limited to epidemiological models, e.g., the compartmental model of (Chamchod and Beier, 2013), which neglects parasite genetics. A population-genetic framework applicable to all human-pathogenic malaria species has to be flexible enough to accommodate super-infections, co-infections, relapses, and recrudescence.

To set up the framework an infection is identified by a vector $\boldsymbol{m} = (m_1, \ldots, m_H)$, where m_h is the number of times haplotype h is infecting. Hence, $m_h = 0$ or $m_h > 0$ if haplotype h is absent or present in the infection, respectively. The number m_h accounts for super-infections with the same haplotype. Moreover, it can be interpreted as the "concentration" of haplotype h if several haplotypes are co-infecting, *etc.*

Let Pr $[\boldsymbol{m}|t]$ be the probability of an infection with configuration $\boldsymbol{m}$ given generation t. The infection might be a new infection or a relapse. The probability of infection $\boldsymbol{m}$, given it occurs in generation t, when the host belongs to group g, and given it is a relapse with a delay of d generations, when the host belonged to group g', is denoted by Pr $[\boldsymbol{m}|t-d, g'; t, g]$. Hence, the probability of infection $\boldsymbol{m}$ occurring in a host in group g in generation t, which is a relapse from generation $t-d$, when the host belonged to group g', is

$$\Pr[\boldsymbol{m}; t-d, g'; t, g] = \Pr[\boldsymbol{m}|t-d, g'; t, g] R_d^{(t)} G_{g',g}^{(t-d,t)}. \qquad (2)$$

The conditional probability Pr $[\boldsymbol{m}|t-d, g'; t, g]$ reflects the model of super- and co-infections. There are many possible models. Super- and co-infections are both notoriously difficult to address. Namely, knowledge about the vector dynamics and the distribution of haplotype combinations in the mosquito population must be known. This is a difficult task and research on the topic is currently expanding, (*cf.* Nkhoma et al., 2012; Wong et al., 2018; Zhu et al., 2019; Nkhoma et al., 2020; Dia and Cheeseman, 2021; Neafsey et al., 2021).

2.3.1 A model for super-infections

Many approaches to estimate MOI or COI by Bayesian or maximum-likelihood methods (e.g. (Hill and Babiker, 1995; Stephens et al., 2001; Rastas et al., 2005; Li et al., 2007; Hastings and Smith, 2008; Druet and Georges, 2010; Ross et al., 2012; Wigger et al., 2013; Taylor et al., 2014; Galinsky et al., 2015; Ken-Dror and Hastings, 2016; Schneider, 2018; Hashemi and Schneider, 2021)) are based on a model, which assumes only super-infections, but no co-infections. The number of super-infections m is referred to as multiplicity of infection (MOI; see Figure 3).

Let $M_m^{(t,g)}$ be the probability that a host belonging to group g is super-infected exactly m times in generation t. This is a probability distribution, hence

$$\sum_{m=1}^{\infty} M_m^{(t,g)} = 1 \tag{3}$$

for all t and g. At each infectious event, exactly one haplotype is randomly drawn from the mosquito population, i.e., the haplotype distribution $\boldsymbol{P}_t$. Hence, given MOI m in generation t, the infection $\boldsymbol{m} = (m_1, \ldots, m_H)$, which indicates how many times haplotype h was transmitted, follows a multinomial distribution with parameters m and $\boldsymbol{P}_t$, i.e.,

$$\Pr[\boldsymbol{m}|m;t] = \binom{m}{\boldsymbol{m}} \boldsymbol{P}_t^{\boldsymbol{m}}, \tag{4}$$

where $\binom{m}{\boldsymbol{m}} := \frac{m!}{\prod_{h=1}^{H} m_h!}$ is a multinomial coefficient, and $\boldsymbol{P}_t^{\boldsymbol{m}} := \prod_{h=1}^{H} P_h^{(t)} m_h$. Clearly, the constraint $|\boldsymbol{m}| := \sum_{h=1}^{H} m_h = m$ must hold. If an infection is a relapse with a delay of d generations, the haplotypes have to be drawn according to the distribution $\boldsymbol{P}_{t-d}$.

Therefore, the probability of infection $\boldsymbol{m}$ given it has MOI $m = |\boldsymbol{m}|$ and occurs in generation t, when the host belongs to group g, from a relapse with a delay of d generations, when the host belonged to group g', is given by

$$\Pr[\boldsymbol{m}, m|t-d, g'; t, g] = M_m^{(t-d,g')} \binom{m}{\boldsymbol{m}} \boldsymbol{P}_{t-d}^{\boldsymbol{m}}, \tag{5}$$

where $M_m^{(t-d,g')}$ is the probability of MOI m in generation $t-d$ of a host in group g'. This model makes the expression (WHO, 2021b) much more explicit.

2.3.2 Choices for the distribution of super-infections

The model (WHO, 2021b) becomes even more explicit for specific choices of the distribution of MOI. A popular choice emerges from the assumption of rare and independent infections, namely that MOI is conditionally Poisson distributed (*cf.* Schneider, 2021), i.e.,

$$M_m^{(t,g)} = \frac{1}{\exp(\lambda_{t,g}) - 1} \frac{\lambda_{t,g}^m}{m!}, \tag{6}$$

where $\lambda_{t,g} > 0$ is the Poisson parameter of group g in generation t and $m = 1, 2, \ldots$.

Another popular choice is the conditional negative-binomial distribution. It is similar to the Poisson distribution but over-dispersed (*cf.* 17).

2.4 The exo-erythrocytic and erythrocytic cycles

Assume an infection subsumed by the vector $\boldsymbol{m}$ having MOI $m = |\boldsymbol{m}|$. Since all steps of reproduction are clonal inside the host, it is not necessary to model the different parasite stages explicitly. Rather, it suffices to model the change in haplotype frequencies inside the host as a single step.

If the host belongs to group g, the 'absolute' frequency of haplotype h is $\frac{m_h}{m} W_{\boldsymbol{m},h}^{(t,g)}$. Here, $W_{\boldsymbol{m},h}^{(t,g)}$ is the fitness in generation t of haplotype h in infection $\boldsymbol{m}$ of a host belonging to group g. It is interpreted as the expected number of gametocyte descendants of a single copy of haplotype h infecting the host at the time a mosquito takes her blood meal.

2.4.1 Host-vector transmission

Concerning host-vector transmission, a mosquito ingests a fraction f of male and female gametocytes at her blood meal. The gametocyte haplotypes ingested are assumed to be proportional to the haplotype frequencies within the host. More precisely, $f \frac{m_h}{m} W_{\boldsymbol{m},h}^{(t,g)}$ male and female haplotype h are ingested from infection $\boldsymbol{m}$ in group g. (Note different fractions f can also be assumed for male and female gametocytes, reflecting an unequal sex ratio.)

2.4.2 Sporogonic cycle

Recombination occurs immediately after the blood meal (see Figure 1), and only parasites descending from the same host can recombine (see Figure 3). Assuming the mosquito bite a host from group g with infection $\boldsymbol{m}$, the probability that a male gamete of haplotype h fertilizes a female i-gamete is the product of their relative frequencies in the mosquito's gut, i.e.,

$$\frac{f \frac{m_h}{m} W_{\boldsymbol{m},h}^{(t,g)}}{f W_{\boldsymbol{m}}^{(t,g)}} \cdot \frac{f \frac{m_i}{m} W_{\boldsymbol{m},i}^{(t,g)}}{f W_{\boldsymbol{m}}^{(t,g)}} = \frac{m_h W_{\boldsymbol{m},h}^{(t,g)} m_i W_{\boldsymbol{m},i}^{(t,g)}}{m^2 W_{\boldsymbol{m}}^{(t,g)} 2}, \tag{7}$$

where

$$f W_{\boldsymbol{m}}^{(t,g)} := f \sum_{j=1}^{H} \frac{m_j}{m} W_{\boldsymbol{m},j}^{(t,g)} \tag{8}$$

is the total amount of parasites in the mosquito's gut. Therefore, the absolute number of such matings is obtained by multiplying the probability of the mating by the total amount of parasites, i.e.,

$$f A_{h,i}^{(t,g)} \tag{9}$$

where

$$A_{\boldsymbol{m},h,i}^{(t,g)} := \frac{m_h W_{\boldsymbol{m},h}^{(t,g)} m_i W_{\boldsymbol{m},i}^{(t,g)}}{m^2 W_{\boldsymbol{m}}^{(t,g)}}. \tag{10}$$

The absolute frequency of haplotype h in the population of mosquitoes, which descends from infections with configuration $\boldsymbol{m}$, given 1) MOI $m = |\boldsymbol{m}|$, 2) the infections occur in generation t, 3) in hosts in group g, which 4) are either novel infections (delay $d = 0$) or relapses with a delay of d generations, is

$$\Pr[\boldsymbol{m}|m; t-d; t, g] \sum_{j,l=1}^{H} f A_{\boldsymbol{m},j,l}^{(t,g)} r(jl \to h), \tag{11}$$

where r ($jl \to h$) is the probability that a mating between gametes with haplotypes j and l lead to offspring of haplotype h.

The absolute number of haplotype h in the mosquito population, which descend from hosts in group g with MOI m, is calculated from the theorem of total probability, i.e., by 'averaging' over all possible infections $\boldsymbol{m}$ with MOI m. Incorporating all relapses it is given by

$$P_h^{*(g,m)}(t) = \sum_{d=0}^{D} R_d^{(t)} \sum_{\boldsymbol{m}:\,|\boldsymbol{m}|=m} \Pr[\boldsymbol{m}|m; t-d; t, g] \times \sum_{j,l=1}^{H} f A_{\boldsymbol{m},j,l}^{(t,g)} r(jl \to h). \tag{12}$$

If an infection in generation t is a relapse from generation $t - d$ the host might have belonged to a different group g' then. Noting, that

$$\Pr[\boldsymbol{m}|t-d; t, g] = \sum_{g'=1}^{S} G_{g',g}^{(t-d,t)} \Pr[\boldsymbol{m}|m; t-d, g'; t, g] \tag{13}$$

equation (Su et al., 2019) can be rewritten as

$$P_h^{*(g,m)}(t) = \sum_{d=0}^{D} R_d^{(t)} \sum_{g'=1}^{S} G_{g',g}^{(t-d,t)} \sum_{\boldsymbol{m}:\,|\boldsymbol{m}|=m} \Pr[\boldsymbol{m}|m; t-d, g'; t, g] \times \sum_{j,l=1}^{H} f A_{\boldsymbol{m},j,l}^{(t,g)} r(jl \to h). \tag{14}$$

2.5 Evolutionary dynamics

To determine the number of haplotypes h in generation t + 1, equation (Ngwa et al., 2016) has to be averaged over all possible groups and values of MOI. Hence, the absolute frequency of haplotype h in the next generation's sporozoite population is

$$P_h^*(t+1) = f \sum_{d=0}^{D} R_d^{(t)} \sum_{g,g'=1}^{S} G_{g',g}^{(t-d,t)} \sum_{m=1}^{\infty} \times \sum_{\boldsymbol{m}:\,|\boldsymbol{m}|=m} \Pr[\boldsymbol{m}, m|t-d, g'; t, g] \times \sum_{j,l=1}^{H} A_{\boldsymbol{m},j,l}^{(t,g)} r(jl \to h). \tag{15}$$

The relative frequency of haplotype h in the sporozoite population in generation t + 1 is hence

$$P_h(t+1) = \frac{P_h^*(t)}{\sum_{i=1}^{H} P_i^*(t)}. \tag{16}$$

The dynamics (Watson et al., 2021) are extremely flexible. They allow to model, e.g., temporal changes in selection pressures (for instance changing treatment policies in the context of drug-resistance evolution, temporally varying transmission intensities, intra-host competition of parasites, super- and co-infections, relapses, recrudescences *etc.*). This however requires to specify the model more explicitly.

Next, we show how this is done if only super-infections but no co-infections are considered.

2.6 Evolutionary dynamics with super-infections

We introduce a couple of simplifying assumptions, which make the model more explicit. First, only super- but no co-infections are assumed. I.e., the super-infection model (Lover et al., 2018) applies and is substituted into (Schneider and Escalante, 2013). Thus, (Schneider and Escalante, 2013), becomes

$$P_h^*(t+1) = f \sum_{d=0}^{D} R_d^{(t)} \sum_{g,g'=1}^{S} G_{g',g}^{(t-d,t)} \sum_{m=1}^{\infty} M_m^{(t-d,g')} \sum_{\boldsymbol{m}:\,|\boldsymbol{m}|=m} \binom{m}{\boldsymbol{m}} \boldsymbol{P}_{t-d}^{\boldsymbol{m}} \times \sum_{j,l=1}^{H} A_{\boldsymbol{m},j,l}^{(t,g)} r(jl \to h). \tag{17}$$

3 Results

The framework is appropriate to investigate numerous evolutionary-genetics aspects in malaria. It would be far too comprehensive to exemplify the full flexibility. Hence, only special cases are illustrated here. We assume that only super-infections but no co-infections occur, i.e., the dynamics (Baird et al., 2018) are assumed. First, we clarify the difference between haplotype frequency and prevalence. Then we focus on a simple model of drug resistance. Although it is applicable to all malaria species, primarily it shall illustrate the differences between *P. falciparum* and *P. vivax*, because there were no reports on drug resistance in any of the other species (Tseha and Tyagi, 2021).

3.1 Frequency and prevalence

The evolutionary genetics of malaria are described as the time-change in the frequency distribution of parasite haplotypes. For instance, monitoring the frequencies of haplotypes, which confer drug resistance is essential. However, concerning the clinical pathogenesis, the occurrence of resistance-conferring haplotypes in infections is more relevant. Due to super- and co-infections the frequency of a haplotype h, i.e., its relative abundance among sporozoites in the mosquito population does not coincide with the probability that haplotype h occurs in an infection. The latter is referred to as the haplotype's prevalence.

If only super-infections are considered, the prevalence of haplotype h in generation t, denoted by $q_h^{(t)}$ is derived in section Prevalence in the Mathematical Appendix. It is given by

$$q_h^{(t)} = 1 - \sum_{d=0}^{D} R_d^{(t)} \sum_{g'=1}^{S} G_{g'}^{(t-d)} U_{g'}^{(t-d)}\left(1 - P_h^{(t-d)}\right), \tag{18}$$

where $U_{g'}^{(t-d)}(x)$ is the probability generating function of the MOI distribution in group g' in generation $t-d$. This function characterizes transmission in group g' in generation $t-d$. From the above expression it is clear that prevalence depends on (i) the frequency of haplotype h, (ii) the distributions of MOI in the various groups, and (iii) the distribution of relapses/recrudescence. If no relapses or recrudescences occur, as it is the case for *P. falciparum* and *P. knowlesi*, the prevalence simplifies to

$$q_h^{(t)} = 1 - \sum_{g=1}^{S} G_g^{(t)} U_g^{(t)}\left(1 - P_h^{(t-d)}\right). \tag{19}$$

Hence, for *P. falciparum* and *P. knowlesi* prevalence is characterized by the haplotype frequency distribution in t, the distribution of groups, and the MOI distributions in the groups. We illustrate the effect of relapses on prevalence in a simple example below.

3.2 Selection at a single locus without intra-host competition

Assume drug resistance is determined by a single locus. This is a reasonable assumption since often drug resistance is determined mainly by mutations at one locus. For instance, in *P. falciparum* resistance to chloroquine is determined by mutations at the *Pfcrt* locus, while resistance artemisinin is determined by mutations in the Kelch-13 propeller region (Cui et al., 2015). The assumption is even justified in sulfadoxine-pyrimethamine resistance, determined by the *Pfdhfr* and *Pfdhps* loci, because mutations at the *Pfdhfr* locus seem to have a much stronger effect (McCollum et al., 2012).

Assume n alleles $A_1, \ldots, A_n$ are segregating at the selected locus. The n different alleles confer different levels of drug resistance. All other alleles are assumed to be neutral. Thus, the number of possible haplotypes, H, is a multiple of n, i.e., $H = nN$. Hence, N is the number of all possible haplotypes when the resistance-conferring locus is disregarded. Let us assume that the haplotypes are ordered such that haplotypes $h = (a-1)N+1, \ldots, aN$ carry allele A_a at the resistance-conferring locus. Therefore, the frequency of allele A_a at time $t+1$, denoted by $p_a^{(t+1)}$ is given by

$$p_a^{(t)} = \sum_{h=(a-1)N+1}^{aN} P_h^{(t)}. \tag{20}$$

Cumulatively, we denote the vector of allele frequencies in generation t by $\boldsymbol{p}_t$.

Under the assumption of no intra-host competition of parasites these dynamics can be made more explicit. In an infection characterized by $\boldsymbol{m}$ of a host in group g, no intra-host competition means that the fitness of an infecting haplotype h is independent of what other haplotypes are present in the infection, i.e., it is independent of $\boldsymbol{m}$, or formally

$$W_{\boldsymbol{m},h}^{(t,g)} = W_h^{(t,g)}. \tag{21}$$

Furthermore, because fitness is only determined by the resistance-conferring locus, the fitness of haplotype h depends only on its allele at this locus. Let the fitness of haplotypes carrying allele A_a at the resistance-conferring locus be denoted by $w_a^{(t,g)}$, i. e,

$$\begin{aligned} w_a^{(t,g)} = W_h^{(t,g)} = W_{\boldsymbol{m},h}^{(t,g)} \quad &\text{for} \quad h \\ = (a-1)N+1, \ldots, aN \quad &\text{and for all} \quad \boldsymbol{m}. \end{aligned} \tag{22}$$

Moreover, let the average fitness of allele A_a in generation t be

$$w_a^{(t)} = \sum_{g=1}^{S} w_a^{(t,g)} G_g^{(t)}. \tag{23}$$

As shown in the Mathematical Appendix the dynamics of the allele frequencies are given by

$$p_a^{(t+1)} = \frac{w_a^{(t)} \sum\limits_{d=0}^{D} R_d^{(t)} p_a^{(t-d)}}{\sum\limits_{b=1}^{n} w_b^{(t)} \sum\limits_{d=0}^{D} R_d^{(t)} p_b^{(t-d)}}. \tag{24}$$

As in the case without relapses/recrudescence (*cf.* 17), these dynamics are independent of the distribution of MOI. This holds because no intra-host competition occurs and because only super-infections are considered. Even without intra-host competition the dynamics of the allele frequencies at the selected locus might depend on MOI, depending on the assumed model for co-infections; a general statement cannot be made.

Further, the dynamics (Bousema et al., 2011) depend only on the average fitnesses of the alleles $w_a^{(t)}$. This implies that the stratification of the host population into different groups does not need to be modelled explicitly, when considering selection at a single locus.

Note that the average fitnesses can be scaled by any constant without affecting the dynamics (Bousema et al., 2011). Hence, it suffices to consider relative fitnesses, and fitness can be normalized such that $w_1^{(t)} = 1$ in every generation.

3.2.1 The effect of recrudescences and relapses

In the dynamics of the allele frequencies (Bousema et al., 2011) the effect of relapses or recrudescence is clearly visible. In the case of no relapses or recrudescence, i.e., $R_d^{(t)} = 0$ for $d \geq 0$ the dynamics simplify to

$$p_a^{(t+1)} = \frac{w_a^{(t)} p_a^{(t)}}{\sum_{b=1}^{n} w_b^{(t)} p_b^{(t)}}. \tag{25}$$

In this situation, the allele frequencies in generation $t + 1$ are solely determined by the fitnesses and the allele frequencies in generation t. Once relapses or recrudescences are considered, the allele frequencies in generation $t + 1$, depend also on the allele frequencies in previous generations. This is intuitively clear, because relapses/recrudescence are equivalent to infections from the sporozoite population from previous generations (see Figure 2). Hence, relapses/recrudescence act as “seed banks”. Intuitively, this will delay the evolutionary dynamics, because the allele frequencies are averaged over several previous generations.

To further discuss the effect of relapses/recrudescence we impose some additional assumptions. First, we assume that the selective environment does not change over time, i.e., $w_a^{(t)} = w_a$ for all t. This is a reasonable assumption when considering drug resistance evolution over a time period in which treatment policies do not change. In this case, the change in allele frequencies can be solved explicitly only in the absence of relapses/recrudescence. Namely, the dynamics become

$$p_a^{(t+1)} = \frac{w_a^{t+1} p_a^{(0)}}{\sum_{b=1}^{n} w_b^{t+1} p_b^{(0)}}, \tag{26}$$

where $p_a^{(0)}$ are the initial allele frequencies in generation $t = 0$. From these dynamics it follows that the average fitnesses w_a can be estimated from longitudinal data of allele frequencies by fitting a straight-line regression (see 48, 17).

Once relapses/recrudescence are considered, the dynamics can no longer be solved explicitly, but need to be calculated recursively from the frequencies of the last $D + 1$ generations, i.e., they become

$$p_a^{(t+1)} = \frac{w_a \sum_{d=0}^{D} R_d^{(t)} p_a^{(t-d)}}{\sum_{b=1}^{n} w_b \sum_{d=0}^{D} R_d^{(t)} p_b^{(t-d)}}. \tag{27}$$

Importantly, to be able to iterate these dynamics, initial frequencies need to be known from D generations in the past. Hence, to calculate the frequencies in generation $t = 1$, initial frequencies $p_a^{(0)}, p_a^{(-1)}, \ldots, p_a^{(-D)}$ need to be specified. Moreover, the distribution $R_d^{(t)}$ needs to be known. In practice, the distribution of relapses might change over time. For instance, changes in control policies impact malaria transmission and hence the proportion of new infection in comparison to relapses. If transmission intensities decrease, relapses amount for a larger fraction of infections. Also the number of transmission cycles during 1 year decrease. Because the distribution of the time to relapse measured in years will not change, the time distribution measured in units of transmission cycles will change. In the simplest case the distribution of relapses remains constant over time, i.e., $R_d^{(t)} = R_d$, the change of allele frequencies is given by

$$p_a^{(t+1)} = \frac{w_a \sum_{d=0}^{D} R_d p_a^{(t-d)}}{\sum_{b=1}^{n} w_b \sum_{d=0}^{D} R_d p_b^{(t-d)}}. \tag{28}$$

Unfortunately, even if the distribution of relapses is constant, the average fitnesses can no longer be estimated by a linear regression.

The distribution of relapses depends crucially on the specific parasite strain (White, 2011). Consider the following example of drug-resistance evolution, with just two alleles: allele A_1 being the drug sensitive wildtype and A_2 the mutant allele conferring drug resistance. The mutant allele first occurs in generation $t = 0$ at frequency $p_2^{(0)} = 0.001$. Let $w_2 = 1 + s$, where s is the selective advantage of the drug resistant allele A_2. We assume $s = 0.1$, i.e., the fitness is increased by 10%, which is strong selection for population-genetic processes, but reasonable for selection for drug-resistance.

Regarding the distribution of relapses, we assume a situation in which 1 year corresponds to 10 transmission cycles. Relapses often occur in periodic patterns (White, 2011). We first assume a pattern which resembles the relapse pattern described by (Hankey et al., 1953) in temperate zones of Korea. Namely, let v be the probability that a malaria episode relapses, i.e., $R_0 = 1 - v$. We assume the first relapse can occur after 10 transmission cycles, and all further relapses after 4 further transmission cycles for a maximum delay of $D = 90$. More precisely, $R_d = \frac{v}{21}$ for $d = 10, 14, 18, 22, \ldots, 90$ and $R_d = 0$ else. As a comparison we assume a simple second pattern of relapses, in which relapses occur 4–50 generations after the initial infection with equal probability, i.e., $R_d = \frac{v}{43}$ for $d = 4, \ldots, 50$. Compared with the first pattern, relapses occur more frequently and earlier.

The evolutionary dynamics are illustrated in Figure 4. Without relapses $v = 0$, the resistance-conferring allele spreads in approximately 110 generations, which corresponds to 11 years, under the assumed number of 10 transmission cycles per year. Relapses substantially slow down the spread of resistance. The

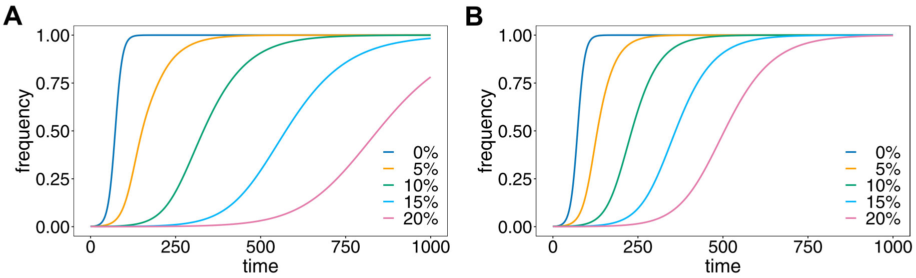


FIGURE 4
Effect of relapses on the evolutionary dynamics. Shown is the frequency of the resistance-conferring allele as a function of time assuming different proportions, *v* of relapses (colors) for the first **(A)** and second **(B)** patterns of relapses.

reason is that relapses act like ‘seed banks’ which retain the frequency distribution of previous generations. For the first pattern (Figure 4A), 5% relapses already substantially delay the spread of resistance to about 400 generations or 40 years. With 20% relapses, the frequency of the mutant allele is just 75% after 1,000 generations corresponding to 100 years. For the second pattern (Figure 4B), the results are qualitatively similar, but relapses have a less profound effect, because they occur with shorter delay after the original infection.

These results provide formal evidence that drug resistance spreads faster in *P. falciparum*, where no relapses occur, than in *P. vivax*, where relapses are common. In fact, while drug resistance is a major concern in *P. falciparum*, it is less common in *P. vivax* (Schneider and Escalante, 2013).

The pattern of relapses depends on 1) genetic factors mediating the frequency of their occurrence; 2) transmission intensities determining the number of malaria generations (transmission cycles per year); 3) the fractions of new infections and relapses; and 4) treatment policies. Particularly, if a drug is partnered with primaquine (PQ) or tafenoquine (TQ) for radical cure, the fraction of relapses reduces, accelerating the spread of resistance to the primary treatment. However, since PQ or TQ also act on gametocytes, they prevent transmission and reduce the selective advantage of drug resistance (*cf.* 23).

3.2.2 Prevalence

Next consider the prevalences corresponding to the evolutionary dynamics illustrated in Figure 4. The evolutionary dynamics are determined by the average fitnesses across the groups of hosts and the distribution of relapses. Consequently, it was not necessary to specify the groups explicitly. However, prevalence given by (Collins and Jeffery, 2007) depends on the generating functions of MOI in the different groups. In the simplest case, which we consider here, the whole population consists of only one group ($S = 1$). Furthermore, we assume that the MOI distribution does not change over time, and follows a conditional Poisson distribution (*cf.* Eq. (6)) with parameter λ. The generating function of this distribution is given by

$$U(x) = \frac{\exp(\lambda x) - 1}{\exp(\lambda x) - 1} \tag{29}$$

(*cf.* 17).

The prevalence of the resistance-conferring allele is obtained from (Collins and Jeffery, 2007) by assuming that haplotypes are characterized by a single locus. Hence,

$$q_2^{(t)} = 1 - \sum_{d=0}^{D} R_d U\left(1 - p_2^{(t-d)}\right) = \sum_{d=0}^{D} R_d \frac{1 - \exp\left(-\lambda p_2^{(t-d)}\right)}{1 - \exp(-\lambda)}. \tag{30}$$

The prevalences corresponding to the dynamics illustrated in Figure 4A, are depicted in Figure 5, assuming different values of the Poisson parameter λ, corresponding to different transmission intensities.

The case $\lambda = 0$, implies that only ‘single-infection’ (one infective event) occurs, in which case prevalence and frequency coincide. As shown in (Schneider, 2021) prevalence always exceeds frequency in the case in which no relapses occur (Figures 5A,F,K). This is intuitive, because the likelihood to observe a parasite variant in an infection increases as the average number of super-infections increase. If transmission intensities are intermediate to high ($\lambda \geq 1$), prevalence is considerably higher than frequency (Figure 5F). If the frequency of the resistance-conferring allele is small, the difference between frequency and prevalence is small in absolute terms, but high in relative terms (compare Figure 5F with Figure 5K).

If relapses occur, the pattern is similar, however, prevalence can be lower than frequency (see Figures 5F–J). The reason is that prevalence is also determined by the frequency distribution of past generations. This occurs only if the average number of super-infections is small (λ slightly larger than 0) and is increasingly pronounced if relapses are more frequent. In general, the difference between frequency and prevalence becomes smaller in absolute and relative terms as the fraction

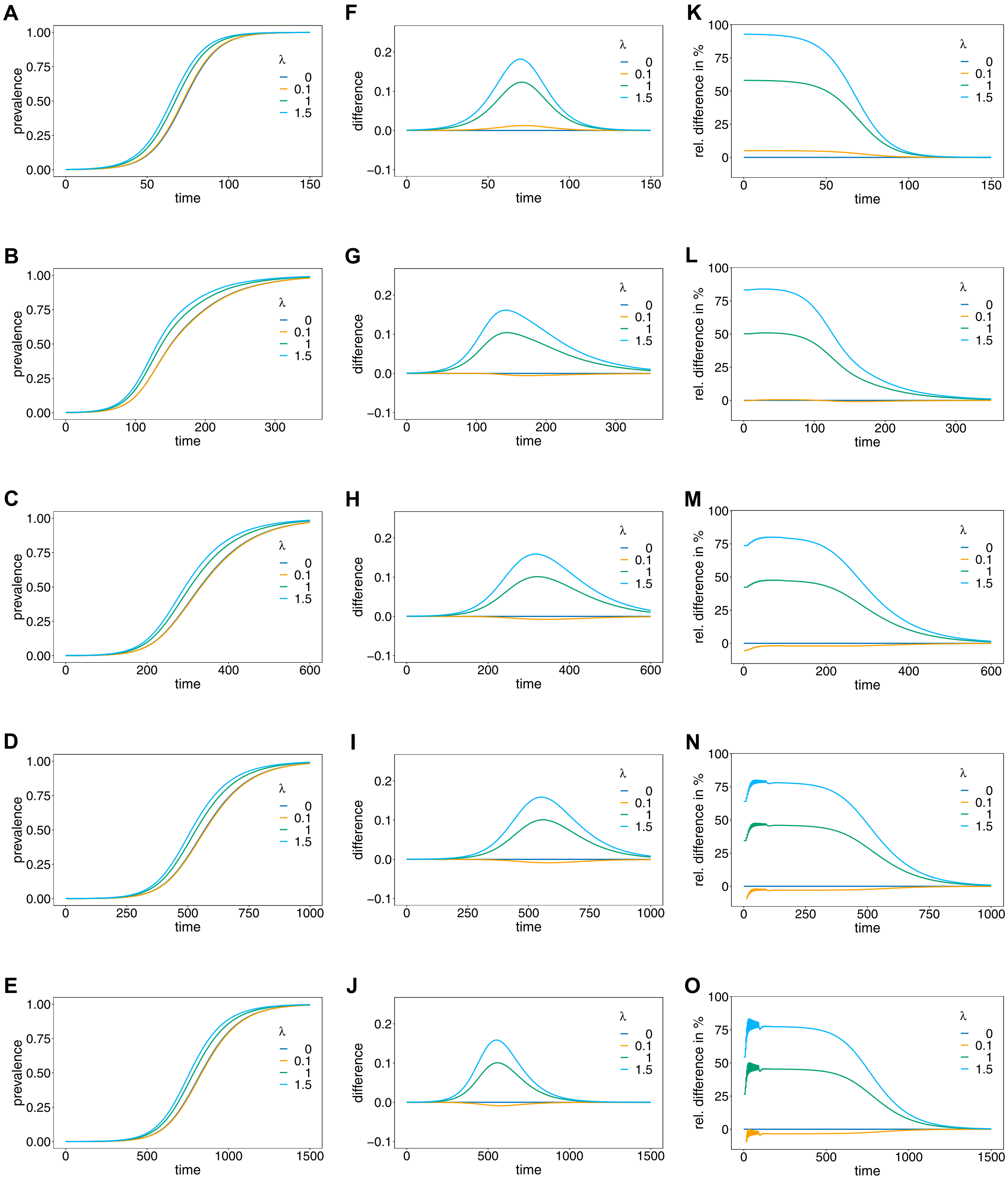


FIGURE 5
Prevalence. Panels **(A–E)** show the prevalence of the resistance-conferring allele corresponding to the dynamics in Figure 4A for different values of the Poisson parameter λ (colors). Panels **(A–E)** correspond to the dynamics with 0%, 5%, 10%, 15%, and 20% relapses, respectively. Panels **(F–J)** show the corresponding difference between prevalence and frequency, and panels **(K–O)** show the corresponding relative difference (prevalence minus frequency divided by frequency) in percent.

of relapses increase. If this fraction is high ($v = 0.15$ or $v = 0.2$) the particular pattern of relapses leads to oscillations in the relative difference between prevalence and frequency, if the frequency of the resistance-conferring allele is low (see Figures 5N,O).

4 Discussion

We introduced a general framework to model evolutionary-genetic processes in malaria, which is flexible enough to capture the characteristics of all human-pathogenic *Plasmodium* species. Such a framework is justified since standard population-genetic theory can only be approximately applied to malaria. The reason is rooted in the malaria transmission cycle, which involves one step of sexual reproduction in the mosquito vectors. A high degree of selfing occurs during this step, because only parasites which descend from the same human host can recombine (*cf.* Figure 3). The framework extends the one introduced in (Schneider and Kim, 2010; Schneider and Kim, 2011; Schneider, 2021), which is only applicable to *P. falciparum*, because it ignores relapses from dormant liver stages as they occur in *P. vivax* and *P. ovale* sp., and recrudescence form prolonged blood stage parasites as they occur in *P. malariae*. These previously widely neglected species are resilient because of relapses and recrudescence, and hence are gaining more importance in the context of malaria eradication. We demonstrated the importance of relapses/recrudescence by contrasting drug resistance-evolution in *P. vivax* and *P. falciparum*.

The necessity to extend the population-genetic framework toward other malaria species is clearly justified by the results presented here. Even in the simplest case of resistance being determined by a single locus, relapses have a profound effect on the evolutionary dynamics, when assuming the same hypothetical drug pressure in both species. Namely, relapses substantially delay the spread of resistance, because they are equivalent—at least in the idealization of the model—to infections with regard to past parasite frequency distributions. In other words, relapses act as seed banks. Dormancy by seed banks is known in evolutionary biology as a bet-hedging strategy that allows organisms to survive through sub-optimal conditions (Shoemaker and Lennon, 2018)—in the case of malaria the absence of the vector. Seed banks are also known to slow down evolutionary processes and influence recombination (Živković and Tellier, 2012; Koopmann et al., 2017; Tellier, 2019). This is no exception in malaria. Although exploring the effect of relapses/recrudescence on recombination was beyond the scope of this work, the effect is rather obvious. Because relapses/recrudescence slow down the evolutionary dynamics, more genetic variation is maintained, leading to a higher level of recombination. In fact, in *P. vivax* higher levels of genetic variations than in *P. falciparum* are a common empirical observation (e.g. (Pacheco et al., 2020)).

Our results have to be understood in a qualitative rather than a quantitative context. Namely, the pattern of relapses have a substantial influence on the evolutionary dynamics. Hence, for adequately predict the spread of resistance, good empirical estimates on the pattern of relapses are necessary. However, empirically distinguishing re-infections (consecutive independent infectious), recrudescence (a rebound of parasitaemia due to incomplete clearance of merozoites), and relapses are notoriously difficult. With more advanced molecular methods becoming available to produce deep-sequencing data (e.g. (Zhong et al., 2018; Gruenberg et al., 2019)), heuristic methods to distinguish recrudescence from reinfections have been proposed (Lin et al., 2015). Also haplotype-based statistical models have been proposed (e.g. (Plucinski et al., 2015)). In principle the framework here can be used to further develop statistical methods to distinguish reinfections from relapses.

To obtain quantitative predictions it is also important to estimate other model parameters. In the context of drug resistance, this includes fitness parameters, metabolic costs for resistance, and the proportion of asymptomatic or untreated infections. The latter can be achieved by routine diagnostics using reliable methods such as ultra-sensitive PCR (e.g. (Gruenberg et al., 2020)). However, also the transmission potential, determined by the abundance of gametocytes has to be determined (*cf.* 9). Selection parameters of drug-resistant haplotypes can be determined from longitudinal molecular data by a linear regressions in *P. falciparum* (McCollum et al., 2012; Schneider, 2021). Disentangling the fitness parameters into metabolic costs and selective advantages of resistance is more difficult. Namely, costs and selective advantages as found in *vitro* studies (*cf.* Cortese and Plowe, 1998) do not linearly scale with *in vivo* observations. In principle, costs can be achieved by contrasting different populations with different drug usage. Comparing such results with *in vitro* studies helps to identify the functional relationship between *in vitro* measurements and *in vivo* observations. Notably, fitness estimates from a linear regression apply mainly to *P. falciparum*. For other malaria species the estimates have to be adapted to the evolutionary dynamics which account for relapses/recrudescence.

Note that the application to modelling drug resistance here had only the purpose of contrasting the absence and presence of relapses. Therefore, only a simplistic model was assumed for drug resistance, i.e., resistance was assumed to be determined by a single biallelic locus. The examples here did not exhibit the full flexibility of the model. If drug resistance occurs in a stepwise fashion as it is found in sulfadoxine-pyrimethamine resistant *P. falciparum* haplotypes (Cortese and Plowe, 1998), where resistance is caused by mutations at several codons in the *Pfdhfr* and *Pfdhps* loci. To capture this situations, resistance-conferring haplotypes have to be modelled by two mulltiallelic loci, where each two-locus haplotype is associated with its own metabolic costs and fitness advantage. Moreover, the mutation haplotypes have to be introduced into the model at different time points. A simple example can be found in (Schneider, 2021).

Relapses are irrelevant in *P. falciparum*, and recrudescences can be neglected, because they occur shortly after the initial infection and do not need to be modeled explicitly. Nevertheless, if transmission intensities are high, which is mainly relevant for *P. falciparum*, the assumption of non-overlapping generations (transmission cycles) are questionable. In the extended framework, relapses can be reinterpreted to mimic overlapping generations. This explains, at least partially, why drug resistance in *P. falciparum* does not necessarily spread first in areas of high transmission (as they occur in Africa) with many more transmission cycles per year.

Reinterpreting relapses in the framework is also important when applied to *P. knowlesi*, which is primarily pathogenic to non-human primates, but became the dominant human-pathogenic malaria species in some endemic areas (Sutherland, 2016). The zoonotic animal-host reservoir renders *P. knowlesi* resilient. Different transmission dynamics between humans and animal hosts can mediate the duration of a transmission cycle. If the number of transmission cycles per year differs among human and non-human primate hosts, this discrepancy can be compensated by modeling overlapping generations by relapses.

We also discussed the differences of frequency and prevalence of parasite haplotypes. The former is the relative abundance of a haplotype in the parasite population, the latter the likelihood that the haplotype occurs in an infection. Studying the haplotype frequency distribution over time is the aim of evolutionary genetics. From a clinical or epidemiological point of view, prevalence is more relevant. The latter is determined by the haplotype frequency distribution and the distribution of super- or co-infections. This was already emphasized in the context of seasonal malaria transmission in (Schneider, 2021) for *P. falciparum*. It was shown that the prevalence of a haplotype always exceeds its frequency. This changes if relapses/recrudescence occur and was exemplified here by the hypothetical dynamics of drug-resistance evolution.

The applications of the framework introduced here are manifold. For instance, in the context of drug resistance, the framework allows to investigate the evolution of multi-drug resistance determined by several loci and changing drug-treatment policies. Also patterns of selection, e.g., genetic hitchhiking, can be studied using this framework. The illustrated applications were only under the simplest assumptions, e.g., of no intra-host competition and super- but no co-infections.

Intra-host competition plays an important role in the spread of HRP2/3 gene deletions associated with false-negative malaria rapid diagnostic tests (RDTs) (Gamboa et al., 2010). Namely, if the treatment guidelines require to verify suspected infections by RDTs before treatment with artemisinin combination therapies (ACTs), as recommended by the WHO (World Health Organization, 2017), false-negative results can lead to delayed treatment. Similarly intra-host competition seems relevant when considering selection on merozoite surface proteins (Goh et al., 2021).

Intra-host dynamics enter the model *via* the definition of fitness. It is not necessary to define an evolutionary-genetic model which captures two timescales, the evolutionary dynamics in terms of generations of transmission cycles, and the timescale of an infectious episode in the same model, as it was done, e.g., in (Kim et al., 2014). Rather, the framework can be used in a multi-scale model, which takes input from a separate intra-host model.

Similarly, the framework does not require to model the mosquito dynamics explicitly. They rather enter *via* the distribution of super- and co-infections. Considering only super-infections has the conceptional advantage, that it is a well-defined model. It is frequently used in statistical approaches to estimate haplotype frequency distributions and MOI (*cf.* e.g. Hill and Babiker, 1995; Stephens et al., 2001; Li et al., 2007; Hastings and Smith, 2008; Wigger et al., 2013; Schneider, 2018; Hashemi and Schneider, 2021). Ignoring co-infections is justified if the distribution of haplotypes in the mosquitoes is uncorrelated or when considering only few loci. However, if one aims to include genetic relatedness, it is important to specify a model for co-infections. This becomes increasingly popular as more high-quality genomic data is becoming available in malaria, which has enough resolution to study genetic relatedness (*cf.* Nkhoma et al., 2012; Wong et al., 2018; Zhu et al., 2019; Nkhoma et al., 2020; Dia and Cheeseman, 2021; Neafsey et al., 2021).

Although the framework is very general, it also has several limitations. First, it ignores mutations. This is not a strong restriction, because in many applications one is interested in *de novo* mutations which occur at discrete time points. This is captured by the model, by introducing new haplotypes (i.e., extending the model) at certain times. However, constant mutation rates, e.g., to study mutation-selection balance, can be easily introduced. Another limitation is the deterministic nature of the framework. When aiming to study stochastic effects such as genetic drift, it is rather straightforward to develop a stochastic version of the framework. Third, the model ignores mitotic recombination during merozoite production inside the host. This plays an important role in some applications, particularly in the structural rearrangement of Var genes (Claessens et al., 2014). These hypervariable genes are responsible to generate important antigen profiles for parasite-host interactions (Warimwe et al., 2009). In any case the framework introduced here allows studying manifold evolutionary-genetic aspects of malaria. Importantly, it allows us to specify benchmark scenarios. More empirical evidence is required to refine relevant parametrizations of the framework.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

KS conceptualized the work, developed the mathematical model, performed the mathematical analysis, produced all figures, and wrote the manuscript. CS assisted to conceptualize the work and wrote the manuscript.

Funding

This study was funded by the Armed Forces Health Surveillance Division (AFHSD), Global Emerging Infections Surveillance (GEIS) Branch, ProMIS ID P0082_22_N6. This work was also supported by grants of the German Academic Exchange (DAAD; https://www.daad.de/de/; Project-ID 57417782, Project-ID: 57599539), the Sächsisches Staatsministerium für Wissenschaft, Kultur und Tourismus and Sächsische Aufbaubank—Förderbank (SMWK-SAB; https://www.smwk.sachsen.de/; https://www.sab.sachsen.de/; project "Innovationsvorhaben zur Profilschärfung an Hochschulen für angewandte Wissenschaften", Project-ID 100257255; project "Innovationsvorhaben zur Profilschärfung 2022", Project-ID: 100613388), the Federal Ministry of Education and Research (BMBF) and the DLR (Project-ID 01DQ20002; https://www.bmbf.de/; https://www.dlr.de/).

Acknowledgments

The authors gratefully acknowledge the constructive comments of the two reviewers and the editor.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1030463/full#supplementary-material

References

Baird, J. K., Battle, K. E., and Howes, R. E. (2018). Primaquine ineligibility in anti-relapse therapy of plasmodium vivax malaria: The problem of g6pd deficiency and cytochrome p-450 2d6 polymorphisms. *Malar. J.* 17, 42–46. doi:10.1186/s12936-018-2190-z

Beshir, K. B., Parr, J. B., Cunningham, J., Cheng, Q., and Rogier, E. (2022). Screening strategies and laboratory assays to support plasmodium falciparum histidine-rich protein deletion surveillance: Where we are and what is needed. *Malar. J.* 21, 201–212. doi:10.1186/s12936-022-04226-2

Bousema, T., Kreuels, B., and Gosling, R. (2011). Adjusting for heterogeneity of malaria transmission in longitudinal studies. *J. Infect. Dis.* 204, 1–3. doi:10.1093/infdis/jir225

Chamchod, F., and Beier, J. C. (2013). Modeling plasmodium vivax: Relapses, treatment, seasonality, and g6pd deficiency. *J. Theor. Biol.* 316, 25–34. doi:10.1016/j.jtbi.2012.08.024

Claessens, A., Hamilton, W. L., Kekre, M., Otto, T. D., Faizullabhoy, A., Rayner, J. C., et al. (2014). Generation of antigenic diversity in plasmodium falciparum by structured rearrangement of var genes during mitosis. *PLoS Genet.* 10, e1004812. doi:10.1371/journal.pgen.1004812

Collins, W. E., and Jeffery, G. M. (2007). Plasmodium malariae: Parasite and disease. *Clin. Microbiol. Rev.* 20, 579–592. doi:10.1128/CMR.00027-07

Cortese, J. F., and Plowe, C. V. (1998). Antifolate resistance due to new and known plasmodium falciparum dihydrofolate reductase mutations expressed in yeast. *Mol. Biochem. Parasitol.* 94, 205–214. doi:10.1016/s0166-6851(98)00075-9

Cui, L., Mharakurwa, S., Ndiaye, D., Rathod, P. K., and Rosenthal, P. J. (2015). Antimalarial drug resistance: Literature review and activities and findings of the icemr network. *Am. J. Trop. Med. Hyg.* 93, 57–68. doi:10.4269/ajtmh.15-0007

Dean, L., and Kane, M. (2020). "Tafenoquine therapy and g6pd genotype," in *Medical genetics summaries*. Editors V. M. Pratt, S. A. Scott, M. Pirmohamed, B. Esquivel, M. S. Kane, and B. L. Kattman (United states: National Center for Biotechnology Information).

Dia, A., and Cheeseman, I. H. (2021). Single-cell genome sequencing of protozoan parasites. *Trends Parasitol.* 37, 803–814. doi:10.1016/j.pt.2021.05.013

Druet, T., and Georges, M. (2010). A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* 184, 789–798. doi:10.1534/genetics.109.108431

Galinsky, K., Valim, C., Salmier, A., de Thoisy, B., Musset, L., Legrand, E., et al. (2015). Coil: A methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malar. J.* 14, 4. doi:10.1186/1475-2875-14-4

Gamboa, D., Ho, M. F., Bendezu, J., Torres, K., Chiodini, P. L., Barnwell, J. W., et al. (2010). A large proportion of p. falciparum isolates in the amazon region of Peru lack pfhrp2 and pfhrp3: Implications for malaria rapid diagnostic tests. *PloS one* 5, e8091. doi:10.1371/journal.pone.0008091

Goh, X. T., Lim, Y. A., Lee, P. C., Nissapatorn, V., and Chua, K. H. (2021). Diversity and natural selection of merozoite surface protein-1 in three species of human malaria parasites: Contribution from south-east Asian isolates. *Mol. Biochem. Parasitol.* 244, 111390. doi:10.1016/j.molbiopara.2021.111390

Gonzales, S. J., Reyes, R. A., Braddom, A. E., Batugedara, G., Bol, S., and Bunnik, E. M. (2020). Naturally acquired humoral immunity against plasmodium falciparum malaria. *Front. Immunol.* 11, 594653. doi:10.3389/fimmu.2020.594653

Greenwood, B., Cairns, M., Chaponda, M., Chico, R. M., Dicko, A., Ouedraogo, J. B., et al. (2021). Combining malaria vaccination with chemoprevention: A promising new approach to malaria control. *Malar. J.* 20, 361–367. doi:10.1186/s12936-021-03888-8

Gruenberg, M., Lerch, A., Beck, H. P., and Felger, I. (2019). Amplicon deep sequencing improves plasmodium falciparum genotyping in clinical trials of antimalarial drugs. *Sci. Rep.* 9, 17790. doi:10.1038/s41598-019-54203-0

Gruenberg, M., Moniz, C. A., Hofmann, N. E., Koepfli, C., Robinson, L. J., Nate, E., et al. (2020). Utility of ultra-sensitive qpcr to detect plasmodium falciparum and plasmodium vivax infections under different transmission intensities. *Malar. J.* 19, 319. doi:10.1186/s12936-020-03374-7

Hankey, D. D., Jones, R., Jr, Coatney, G. R., Alving, A. S., Coker, W. G., Garrison, P. L., et al. (1953). Korean vivax malaria. i. natural history and response to chloroquine. *Am. J. Trop. Med. Hyg.* 2, 958–969. doi:10.4269/ajtmh.1953.2.958

Hashemi, M., and Schneider, K. A. (2021). Bias-corrected maximum-likelihood estimation of multiplicity of infection and lineage frequencies. *PloS one* 16, e0261889. doi:10.1371/journal.pone.0261889

Hastings, I. M., and Smith, T. A. (2008). MalHaploFreq: A computer programme for estimating malaria haplotype frequencies from blood samples. *Malar. J.* 7, 130. doi:10.1186/1475-2875-7-130

Hedrick, P. W. (2011). Population genetics of malaria resistance in humans. *Heredity* 107, 283–304. doi:10.1038/hdy.2011.16

Hill, W. G., and Babiker, H. A. (1995). Estimation of numbers of malaria clones in blood samples. *Proc. Biol. Sci.* 262, 249–257. doi:10.1098/rspb.1995.0203

Ken-Dror, G., and Hastings, I. M. (2016). Markov chain Monte Carlo and expectation maximization approaches for estimation of haplotype frequencies for multiply infected human blood samples. *Malar. J.* 15, 430. doi:10.1186/s12936-016-1473-5

Kim, Y., Escalante, A. A., and Schneider, K. A. (2014). A population genetic model for the initial spread of partially resistant malaria parasites under anti-malarial combination therapy and weak intrahost competition. *PLOS ONE* 9, e101601–e101615. doi:10.1371/journal.pone.0101601

Koopmann, B., Müller, J., Tellier, A., and Živković, D. (2017). Fisher–wright model with deterministic seed bank and selection. *Theor. Popul. Biol.* 114, 29–39. doi:10.1016/j.tpb.2016.11.005

Li, X., Foulkes, A. S., Yucel, R. M., and Rich, S. M. (2007). An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. *Stat. Appl. Genet. Mol. Biol.* 6, 33. doi:10.2202/1544-6115.1321

Lin, J. T., Hathaway, N. J., Saunders, D. L., Lon, C., Balasubramanian, S., Kharabora, O., et al. (2015). Using amplicon deep sequencing to detect genetic signatures of plasmodium vivax relapse. *J. Infect. Dis.* 212, 999–1008. doi:10.1093/infdis/jiv142

Lover, A. A., Baird, J. K., Gosling, R., and Price, R. N. (2018). Malaria elimination: Time to target all species. *Am. J. Trop. Med. Hyg.* 99, 17–23. doi:10.4269/ajtmh.17-0869

McCollum, A. M., Schneider, K. A., Griffing, S. M., Zhou, Z., Kariuki, S., Ter-Kuile, F., et al. (2012). Differences in selective pressure on dhps and dhfr drug resistant mutations in Western Kenya. *Malar. J.* 11, 77. doi:10.1186/1475-2875-11-77

Ndiaye, Y. D., Hartl, D. L., McGregor, D., Badiane, A., Fall, F. B., Daniels, R. F., et al. (2021). Genetic surveillance for monitoring the impact of drug use on plasmodium falciparum populations. *Int. J. Parasitol. Drugs Drug Resist.* 17, 12–22. doi:10.1016/j.ijpddr.2021.07.004

Neafsey, D. E., Taylor, A. R., and MacInnis, B. L. (2021). Advances and opportunities in malaria population genomics. *Nat. Rev. Genet.* 22, 502–517. doi:10.1038/s41576-021-00349-5

Ngwa, C. J., Rosa, T., and Pradel, G. (2016). *The biology of malaria gametocytes*. Rijeka, Croatia: IntechOpen.

Nicoletti, M. (2020). Three scenarios in insect-borne diseases. *Insect-Borne Dis. 21st Century* 2020, 99–251. doi:10.1016/b978-0-12-818706-7.00005-x

Nkhoma, S. C., Nair, S., Cheeseman, I. H., Rohr-Allegrini, C., Singlam, S., Nosten, F., et al. (2012). Close kinship within multiple-genotype malaria parasite infections. *Proc. Biol. Sci.* 279, 2589–2598. doi:10.1098/rspb.2012.0113

Nkhoma, S. C., Trevino, S. G., Gorena, K. M., Nair, S., Khoswe, S., Jett, C., et al. (2020). Co-Transmission of related malaria parasite lineages shapes within-host parasite diversity. *Cell Host Microbe* 27, 93–103. e4. doi:10.1016/j.chom.2019.12.001

Pacheco, M. A., Forero-Peña, D. A., Schneider, K. A., Chavero, M., Gamardo, A., Figuera, L., et al. (2020). Malaria in Venezuela: Changes in the complexity of infection reflects the increment in transmission intensity. *Malar. J.* 19, 176. doi:10.1186/s12936-020-03247-z

Plucinski, M. M., Morton, L., Bushman, M., Dimbu, P. R., and Udhayakumar, V. (2015). Robust algorithm for systematic classification of malaria late treatment failures as recrudescence or reinfection using microsatellite genotyping. *Antimicrob. Agents Chemother.* 59, 6096–6100. doi:10.1128/AAC.00072-15

Rastas, P., Koivisto, M., Mannila, H., and Ukkonen, E. (2005). "A hidden markov technique for haplotype reconstruction," in *Algorithms in bioinformatics*. Editors R. Casadio and G. Myers (Berlin, Heidelberg: Springer), 140–151. Lecture Notes in Computer Science. doi:10.1007/11557067_12

Ross, A., Koepfli, C., Li, X., Schoepflin, S., Siba, P., Mueller, I., et al. (2012). Estimating the numbers of malaria infections in blood samples using high-resolution genotyping data. *Plos One* 7, e42496. doi:10.1371/journal.pone.0042496

Schneider, K. A. (2021). *Charles Darwin meets ronald Ross: A population-genetic framework for the evolutionary dynamics of malaria*. (Cham: Springer International Publishing), 149–191. chap. 6. doi:10.1007/978-3-030-50826-5_6

Schneider, K. A., and Escalante, A. A. (2013). Fitness components and natural selection: Why are there different patterns on the emergence of drug resistance in plasmodium falciparum and plasmodium vivax? *Malar. J.* 12, 15–11. doi:10.1186/1475-2875-12-15

Schneider, K. A., and Kim, Y. (2010). An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance. *Theor. Popul. Biol.* 78, 93–108. doi:10.1016/j.tpb.2010.06.005

Schneider, K. A., and Kim, Y. (2011). Approximations for the hitchhiking effect caused by the evolution of antimalarial-drug resistance. *J. Math. Biol.* 62, 789–832. doi:10.1007/s00285-010-0353-9

Schneider, K. A. (2018). Large and finite sample properties of a maximum-likelihood estimator for multiplicity of infection. *PloS one* 13, e0194148. doi:10.1371/journal.pone.0194148

Schneider, K. A., Tsoungui Obama, H. C. J., Kamanga, G., Kayanula, L., and Adil Mahmoud Yousif, N. (2022). The many definitions of multiplicity of infection. *Front. Epidemiol.* 2, 961593. doi:10.3389/fepid.2022.961593

Selvaraj, P., Wenger, E. A., and Gerardin, J. (2018). Seasonality and heterogeneity of malaria transmission determine success of interventions in high-endemic settings: A modeling study. *BMC Infect. Dis.* 18, 413–414. doi:10.1186/s12879-018-3319-y

Shoemaker, W. R., and Lennon, J. T. (2018). Evolution with a seed bank: The population genetic consequences of microbial dormancy. *Evol. Appl.* 11, 60–75. doi:10.1111/eva.12557

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989. doi:10.1086/319501

Su, Xz, Lane, K. D., Xia, L., Sá, J. M., and Wellems, T. E. (2019). Plasmodium genomics and genetics: New insights into malaria pathogenesis, drug resistance, epidemiology, and evolution. *Clin. Microbiol. Rev.* 32, e00019. doi:10.1128/CMR.00019-19

Sutherland, C. J. (2016). Persistent parasitism: The adaptive biology of malariae and ovale malaria. *Trends Parasitol.* 32, 808–819. doi:10.1016/j.pt.2016.07.001

Taylor, A. R., Flegg, J. A., Nsobya, S. L., Yeka, A., Kamya, M. R., Rosenthal, P. J., et al. (2014). Estimation of malaria haplotype and genotype frequencies: A statistical approach to overcome the challenge associated with multiclonal infections. *Malar. J.* 13, 102. doi:10.1186/1475-2875-13-102

Tellier, A. (2019). Persistent seed banking as eco-evolutionary determinant of plant nucleotide diversity: Novel population genetics insights. *New Phytol.* 221, 725–730. doi:10.1111/nph.15424

Tseha, S. T. (2021). "Plasmodium species and drug resistance," in *Plasmodium species and drug resistance*. Editor R. K. Tyagi (Rijeka: IntechOpen). chap. 2. doi:10.5772/intechopen.98344

Warimwe, G. M., Keane, T. M., Fegan, G., Musyoki, J. N., Newton, C. R., Pain, A., et al. (2009). Plasmodium falciparum var gene expression is modified by host immunity. *Proc. Natl. Acad. Sci. U. S. A.* 106, 21801–21806. doi:10.1073/pnas.0907590106

Watson, J. A., Nekkab, N., and White, M. (2021). Tafenoquine for the prevention of plasmodium vivax malaria relapse. *Lancet. Microbe* 2, e175–e176. doi:10.1016/S2666-5247(21)00062-8

White, N. J. (2011). Determinants of relapse periodicity in plasmodium vivax malaria. *Malar. J.* 10, 297. doi:10.1186/1475-2875-10-297

WHO (2021). *Global technical strategy for malaria 2016–2030*. Geneva, Switzerland: World Health Organization.

WHO (2021). World malaria report 2020: 20 years of global progress and challenges. Available at: www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2020.

Wigger, L., Vogt, J. E., and Roth, V. (2013). Malaria haplotype frequency estimation. *Stat. Med.* 32, 3737–3751. doi:10.1002/sim.5792

Wong, W., Wenger, E. A., Hartl, D. L., and Wirth, D. F. (2018). Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission. *PLoS Comput. Biol.* 14, e1005923. doi:10.1371/journal.pcbi.1005923

World Health Organization, (2017) A framework for malaria elimination.

Zhong, D., Lo, E., Wang, X., Yewhalaw, D., Zhou, G., Atieli, H. E., et al. (2018). Multiplicity and molecular epidemiology of plasmodium vivax and plasmodium falciparum infections in east Africa. *Malar. J.* 17, 185. doi:10.1186/s12936-018-2337-y

Zhu, S. J., Hendry, J. A., Almagro-Garcia, J., Pearson, R. D., Amato, R., Miles, A., et al. (2019). The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. *eLife* 8, e40845. doi:10.7554/eLife.40845

Živković, D., and Tellier, A. (2012). Germ banks affect the inference of past demographic events. *Mol. Ecol.* 21, 5434–5446. doi:10.1111/mec.12039

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 29 November 2022
DOI 10.3389/fgene.2022.1060898

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU), United States

REVIEWED BY
Zhong-duo Wang,
Guangdong Ocean University, China
Martijn Derks,
Wageningen University and Research, Netherlands

*CORRESPONDENCE
Željka Pezer,
zpezer@irb.hr

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

RECEIVED 03 October 2022
ACCEPTED 15 November 2022
PUBLISHED 29 November 2022

CITATION
Pokrovac I and Pezer Ž (2022), Recent advances and current challenges in population genomics of structural variation in animals and plants.
Front. Genet. 13:1060898.
doi: 10.3389/fgene.2022.1060898

Recent advances and current challenges in population genomics of structural variation in animals and plants

Ivan Pokrovac and Željka Pezer*

Laboratory for Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

The field of population genomics has seen a surge of studies on genomic structural variation over the past two decades. These studies witnessed that structural variation is taxonomically ubiquitous and represent a dominant form of genetic variation within species. Recent advances in technology, especially the development of long-read sequencing platforms, have enabled the discovery of structural variants (SVs) in previously inaccessible genomic regions which unlocked additional structural variation for population studies and revealed that more SVs contribute to evolution than previously perceived. An increasing number of studies suggest that SVs of all types and sizes may have a large effect on phenotype and consequently major impact on rapid adaptation, population divergence, and speciation. However, the functional effect of the vast majority of SVs is unknown and the field generally lacks evidence on the phenotypic consequences of most SVs that are suggested to have adaptive potential. Non-human genomes are heavily under-represented in population-scale studies of SVs. We argue that more research on other species is needed to objectively estimate the contribution of SVs to evolution. We discuss technical challenges associated with SV detection and outline the most recent advances towards more representative reference genomes, which opens a new era in population-scale studies of structural variation.

KEYWORDS

genomic structural variation, structural variants, copy number variation, population genomics, evolution, adaptation, speciation

Changing definition and typical properties

Genetic variation is the major focus of population genetics as it provides raw material upon which evolutionary forces act to create phenotypic diversity. Over the past two decades, it has become evident that variation in the linear structure of the genome is taxonomically ubiquitous and that it affects a much larger portion of the genome than the variation in the sequence itself (Huddleston et al., 2017; Kosugi et al., 2019; Hämälä et al., 2021; Box 1). This form of genetic variation results in structural variants (SVs) that can affect orientation (inversions), position (translocations), or copy number. The latter are collectively termed copy number variants (CNVs) and include deletions, insertions, and amplifications of a sequence. A specific group of CNVs termed presence-absence variations (PAVs) refers to sequences that exist in some genomes while completely missing in other genomes of the same species (Saxena et al., 2014). SVs were first defined as events of at least 1 kilobase pairs (kbp) in length (Feuk et al., 2006) but the definition has since expanded to encompass sizes down to 50 bp and larger (Alkan et al., 2011; Sudmant et al., 2015b). Our increasing understanding of the prevalence of SVs as major contributors to genetic variation has led to the inclusion of other genome rearrangements and elements in this definition, that were long before known to have a variable structure within population. The current definition based on SV size also includes interspersed elements (such as transposable elements; TEs), tandem repeats (including micro-, mini-, and macrosatellites) as well as aneusomy and aneuploidy (Pös et al., 2021).

Spontaneous, *de novo* SVs occur several hundred-fold less frequently than point mutations (Belyeu et al., 2021), although the mutation rate varies considerably by SV type (Collins et al., 2020). Recent large family-trio studies in humans and rhesus monkeys estimated that less than one *de novo* CNV is formed per genome per generation (Belyeu et al., 2021; Thomas et al., 2021). Interestingly, parental age does not affect the rate of these mutations in either species, in contrast to single nucleotide variants (SNVs), which accumulate with paternal age in both species (Kong et al., 2012; Wang et al., 2020). This difference between SNVs and CNVs was proposed to be due to the mechanism of their formation—CNVs are thought to form during meiosis which occurs only once per generation, whereas SNVs can arise as errors during replication in mitosis or unrepaired DNA damage—processes which occur frequently over a lifetime in the germline (Thomas et al., 2021).

Some genomic regions show an extraordinary propensity for structural variation such that they reach mutation rates hundreds and thousands of times higher than nucleotide substitutions, according to some estimates (Zhang et al., 2009). These are referred to as recurrent SVs. Their high mutability is attributable to the repetitive architecture of the genomic region in which they reside, which enables non-allelic homologous recombination (NAHR). Among all known mechanisms of SV formation, NAHR is thought to occur the most frequently, when two highly similar but non-allelic DNA sequence repeats align and crossover during meiosis, causing deletion, duplication, or inversion of the region between the repeats, depending on the orientation of the aligned sequences (Zhang et al., 2009). These mediators of NAHR are usually considered to be CNVs themselves as they exist in the genome in variable low or high copy numbers, such as segmental duplications, transposable elements, and tandem repeats. Other mechanisms of SV formation such as non-homologous end joining (NHEJ), microhomology-mediated break-induced replication (MMBIR), fork stalling and template switching (FoSTeS), and replication slippage are not dependent on high sequence similarity and create mainly non-recurrent SVs. These mechanisms and events are usually discussed in the context of genomic disorders (Hastings et al., 2009; Carvalho and Lupski, 2016), although they may contribute to natural polymorphism without seemingly negative effects.

Effect on gene expression and phenotypic variation

The high mutability of SVs is reflected in their high variability within population. For example, it is currently estimated that any human individual contains on average 16 Mb of structural variation (Ebert et al., 2021) or up to 27,000 SVs, including highly repetitive elements (Chaisson et al., 2019; see Box 1). According to the data from NCBI's database of human genomic structural variation (dbVar), almost 100,000 regions in the human genome are affected by SVs at population frequency ≥1% (Box 1). Given this abundance and high variability within population, SVs are expected to have a large impact on phenotypic variation. However, determining the functional effects of the majority of SVs is difficult, especially in natural populations which are not readily amenable to genetic manipulations (Lauer and Gresham 2019). The association of SVs with gene expression remains the most commonly used proxy for assigning phenotypic consequences. An ever-increasing number of population-scale studies have emerged to suggest that SVs of all types contribute to phenotypic variation on multiple layers of gene regulation. CNVs can alter gene dosage (Handsaker et al., 2015) and thus directly affect protein levels, as shown for the human salivary amylase gene (Perry et al., 2007). Structural variants can also modulate gene expression by re-organizing chromatin domains. Perturbations of topologically associated domains (TADs) can lead to the formation of novel regulatory modules, as shown in humans, apes, and mice (Spielmann et al., 2018; Fudenberg and Pollard, 2019; Gilbertson et al., 2022). CNVs can encompass regulatory elements, such as in the case of an enhancer that controls a gene *NDP* that is responsible for wing pigmentation in pigeons (Vickrey et al., 2018). Expression of this gene is positively correlated with both increased melanism and enhancer copy number. In crows, the same gene is associated with plumage variation but is controlled by a different SV type - an LTR retrotransposon insertion that causes reduced expression (Weissensteiner et al., 2020). SVs can affect whole regulatory networks by affecting single key transcription factors and thus have a large phenotypic effect. This was recently exemplified by a mutation in the *ENO* gene, which encodes a transcription factor that regulates floral meristem size in tomatoes - an 85-bp deletion in the promoter of *ENO* was shown to be responsible for the increase in fruit size during tomato domestication (Yuste-Lisbona et al., 2020). Copy number variation in introns causes variable gene length and is commonly found in healthy human populations. These CNVs reside inside genes with essential functions and are proposed to be responsible for their differential regulation between individuals (Rigau et al., 2019). A recent genome-wide association study (GWAS) based on presence-absence variations in rapeseed identified PAVs among different ecotypes that altered the expression of genes responsible for flowering regulation (Song et al., 2020).

While these and other studies illustrate the contribution of individual SVs to phenotypic variation *via* gene regulation, they do not attest to the extent to which SVs explain overall variation in gene transcription within population. Several studies to date have attempted to ascertain the causality of SVs at expression quantitative trait loci (eQTLs). The most comprehensive study thus far, performed in humans and based on over 600 individuals and 48 tissues, found that SVs are causal at 2.66% of eQTLs which represents a tenfold enrichment relative to their abundance in the genome (Scott et al., 2021). This study revealed that, among all SV types, multiallelic CNVs, both coding and non-coding, have the highest association with eQTLs and that the contribution of transposable element insertions was small. Prior estimates based on a limited number of samples and tissues are in discordance with the study by Scott et al. (2021), as they found either a much larger or much smaller proportion of eQTLs to be caused by SVs. For example, a study based on 13 tissues from 147 individuals estimated up to 6.8% of eQTLs are driven by a causal SV (Chiang et al., 2017). An earlier study associated only 0.56% of eQTLs with SVs (Sudmant et al., 2015b), but it was based on a single cell line although the number of individuals was comparable to the study performed by Scott et al. (2021). This large disagreement in estimates between studies suggests that future efforts should employ a more exhaustive number of tissue types, and possibly target a variety of biological processes, to more precisely assess the contribution of SVs on gene expression in a tissue- and condition-specific manner. Indeed, genes with tissue-specific expression exhibit greater copy number variability than genes with widespread expression (Dopman and Hartl, 2007; Henrichsen et al., 2009; Keel et al., 2016), suggesting that SVs more often have roles in specialized rather than general processes. A recent study based on only two tissue types in three-spined sticklebacks found a strong positive correlation between gene copy number and expression in almost 40% of analyzed CNVs (Huang et al., 2019). Such high association becomes less surprising when one considers that gene-encompassing CNVs were previously found to be enriched for immune activity genes in sticklebacks and that the study focused on immune tissues where these genes are expected to be expressed. Another study identified thousands of tandemly repeated minisatellite sequences variable in copy number within population to be associated with local expression and DNA methylation levels (Garg et al., 2021). These CNVs were associated with genes that have been linked with human phenotypes through genome-wide association studies and were strongly enriched for regulatory elements such as enhancers and promoters, suggesting that these non-coding multiallelic CNVs may be causal for human phenotypes and have regulatory functions.

In summary, multiallelic CNVs seem to be a class of SVs that is the most strongly implicated in the contribution of SVs to variation in gene expression. However, the presented figures are likely underestimates. We can expect to approach more precise estimates with the addition of a more comprehensive set of tissues and by analyzing diverse biological conditions in future studies. Despite the large discrepancies in estimates, current knowledge collectively suggests that both coding and non-coding SVs may have a tremendous impact on gene expression, and thus affect phenotypes in the ways we are just beginning to understand. GWASs based on SNVs have not been able to completely identify the genetic components underpinning (human) traits and disorders; over the past decade, a growing body of evidence has accumulated to suggest SVs as a source of this “missing heritability” (Sudmant et al., 2015b; De Coster et al., 2021; Garg et al., 2022; Zhou et al., 2022).

Impact on evolution

Hundreds of CNVs can be found in the genomes of healthy individuals and they show strong signatures of population structure in numerous species (Sudmant et al., 2015a; Pezer et al., 2015; Xu et al., 2016). This has been used as an argument to propose that the majority of CNVs evolve under neutral evolutionary pressures, such that the patterns of copy number variation seen in populations are mainly shaped by demographic events, mutation rate, and genetic drift (Iskow et al., 2012). However, even such generalizations of the evolutionary implications of CNVs (and other SVs) should be considered in their functional contexts. Recent studies in humans and rhesus monkeys revealed that *de novo* gene deletions outnumber duplications by several times (Belyeu et al., 2021; Thomas et al., 2021), but this ratio becomes skewed over time, as illustrated by the proportion of fixed gene losses along the primate lineage, which becomes smaller (Fortna et al., 2004; Dumas et al., 2007; Sudmant et al., 2013; Thomas et al., 2021). This suggests that, over generations, purifying selection acts against deletions of complete genes. The vast majority of SVs seem to be depleted from functional regions of the genome and segregate at low frequencies, as shown by studies in different species (Pezer et al., 2015; Hämälä et al., 2021). Signals of pervasive selection against all types of SVs that overlap genes, except whole-gene duplications, have recently been discovered in a large analysis of thousands of human genomes (Collins et al., 2020). These studies collectively suggest that most SVs affecting genes are deleterious. A somewhat contrasting observation came from a recent study that suggested that SVs significantly contribute to non-neutral variation in humans (Saitou et al., 2022). Assuming that majority of SVs evolve neutrally, this study looked for SVs with unusual allele frequency distribution among populations and came to a surprising number of over 500 putatively adaptive SVs in humans. A proportion of these included SVs that affect exons and were dominated by multiallelic CNVs.

Contribution to adaptation

Structural variants exist in extremely heterogeneous forms, in terms of type (insertion, deletion, duplication, inversion, and translocation), size, mutation rate, and genomic context. Consequently, even without technical difficulties in their discovery, they constitute a substantial challenge for evolutionary studies. While the current picture of the evolutionary effects of SVs remains incomplete, their contribution to adaptive evolution and diversification is becoming more evident (Radke and Lee, 2015; Saitou et al., 2022). An increasing number of studies suggest that SVs are involved in a variety of adaptations in a range of taxonomic groups, affecting different biological systems such as immunity, metabolism, and sensory perception. Instances of naturally occurring parallel ecological divergence provide an especially useful framework for detecting potentially adaptive SVs. The idea is that if the frequency of an SV is higher in a derived population of a certain ecotype compared to the ancestral population of a different ecotype, and this is observed repeatedly in multiple independent populations, that SV is likely contributing to the adaptive phenotype. Adaptation of marine fish to freshwater represents such a system. A study by Ishikawa et al. (2019) found that a gene involved in fatty acid desaturation was duplicated in freshwater lineages. Transgenic manipulation of this gene enabled marine lineages to produce fatty acids and survive in freshwater that lacks fatty acids. This suggested that differences in gene dosage contribute to differences in survival on fatty acid–deficient diets. In a follow-up study, additional gene duplications were identified to be associated with freshwater colonization, including genes involved in immune function and thyroid hormone metabolism (Ishikawa et al., 2022). In another study, two large chromosome inversions were identified to exhibit parallel association with freshwater adaptation (Zong et al., 2021). These inversions contained multiple genes involved in various processes such as metabolism, immunoregulation, growth, maturation, and osmoregulation, thus potentially affecting morphology, physiology and behavior. It was recently found that large inversions were common and widespread in natural populations of deer mice and several inversions with significant differences in allele frequency between forest and prairie ecotypes were identified, which likely contribute to local adaptation (Harringmeyer and Hoekstra, 2022). It has been proposed that among all SVs, chromosomal inversions are the most frequently linked to adaptive traits (reviewed in Wellenreuther and Bernatchez 2018). However, a wealth of studies suggests that CNVs may be comparable if not even dominant in this aspect. Since the initial discovery of copy number variation, more and more instances of CNVs with a putative role in local adaptation of human populations are emerging (Iskow et al., 2012; Hsieh et al., 2019; Quan et al., 2021; Saitou et al., 2022). Both deletions and duplications are implicated. For example, recurring exonic deletions in the haptoglobin gene were shown to contribute to human health by lowering cholesterol levels in the blood (Boettger et al., 2016). Copy number variations in genes *Ppd-B1* and *Vrn-A1* contribute to global adaptation of wheat to a wide range of environmental conditions (Würschum et al., 2015). These genes modulate the timing of flowering and their increase in copy number is associated with altered expression (Díaz et al., 2012). Furthermore, an increase in *EPSPS* gene copy number confers resistance to the herbicide glyphosate in different weed species (Gaines et al., 2010; Baek et al., 2021). Similarly, triplication of a gene associated with aluminum tolerance in some maize lines correlates with increased expression, which confers higher tolerance to aluminum in maize grown on acidic soils (Maron et al., 2013). Huang et al. (2019) studied the role of gene copy number in adaptation to distinct parasite environments between the lake and river habitats in sticklebacks. In some of these genes, copy number was differentiated between ecotypes and it positively correlated with transcript level, suggesting that gene dosage contributes to local adaptation by modulating expression. Similarly, specific SVs with signs of local adaptation were recently uncovered in chocolate tree, some of which are linked to genes that are also differentially expressed between populations (Hämälä et al., 2021). They were enriched for functions related to immunity, emphasizing the role of SVs in local adaptation to specific pathogens. In the fruit fly, hundreds of TEs were identified to be associated with expression variation of nearby genes, some of them bearing adaptive signatures (Rech et al., 2022). Gene loss can also produce adaptive phenotypes, as suggested for polar bear evolution, where a considerable number of genes encoding olfactory receptors have been lost, as well as the salivary amylase-encoding gene and genes involved in fatty acid metabolism (Rinker et al., 2019). These CNVs evolved rapidly over a short evolutionary period, driven by a dietary shift from omnivorous to carnivorous during polar bear evolution. Even some gene retrocopies show signatures of positive selection, as shown by recent studies in humans and mice (Schrider et al., 2013; Zhang and Tautz 2022).

Contribution to speciation

Changes in genome structure can lead to incompatibilities between populations and thus enhance speciation. SVs can enable reproductive isolation through various mechanisms such as suppressed recombination, hybrid incompatibility, and intrinsic postzygotic or premating isolation (reviewed in Zhang et al., 2021). Inversions, especially large ones, seem to be particularly implicated in suppressing recombination. In heterozygotes for such SV, inverted region fails to pair with non-inverted allele during meiosis, preventing them from cross-over. This results in both variants independently accumulating mutations in their sequences over time, creating "genomic islands of divergence," which eventually leads to incompatibilities (Zong et al., 2021). Incompatibility can also be caused by CNVs, especially if affecting the whole gene, as exemplified by a duplication of a key photosynthetic gene in the yellow monkeyflower (Zuellig and Sweigart, 2018). This variant causes lethality in naturally occurring hybrids between two closely related species, presumably by misregulated transcription. Copy number variation can also play a role in assortative mate choice, as suggested for hundreds of CNVs found to be associated with reinforcement of sexual isolation between the two European subspecies of the house mouse (North et al., 2020). Premating isolation can even be mediated by TE, as shown for the 2.25-kb LTR retrotransposon insertion which affects plumage in birds, a trait associated with prezygotic isolation through social and sexual selection (Weissensteiner et al., 2020). This study nicely illustrates how even a single and small SV can change the evolutionary trajectory of a population and potentially lead to divergence and speciation. Translocations represent a special type of SVs, in that they are often associated with genome instability and negative outcome such as infertility and oncogenesis (Aplan, 2006; Mitelman et al., 2007). Rare instances are found as naturally occurring polymorphisms in healthy individuals. One of the well-studied examples is Robertsonian fusion in the house mouse subspecies *Mus musculus domesticus*, which refers to the translocation of the whole chromosome arm, *i.e.* the joining of two telocentric chromosomes to create a metacentric chromosome. Robertsonian fusions are more frequent in small and geographically isolated populations and they are proposed to contribute to reproductive isolation (Garagna et al., 2014). Similar to inversions, translocations have been associated with the suppression of recombination and have recently been implicated in genetic divergence between subspecies of bananas (Martin et al., 2020) and populations of spiny frogs (Xia et al., 2020).

Structural variants as loci of large effect

Some SVs are large enough to span many genes and regulatory regions. Consequently, they can simultaneously affect multiple traits, acting as supergenes of large effect. Such a role has often been assigned to large inversions, proposed to be associated with complex phenotypes (reviewed in Wellenreuther and Bernatchez 2018). An inversion that contains multiple advantageous alleles will be more strongly selected for than an inversion containing a single favorable gene variant. These alleles are also more likely to be coinherited due to suppressed recombination in heterokaryotypes, contributing further to their rapid increase in frequency in the population under selection. Consequently, large inversions are considered to have significant roles in rapid environmental adaptation and speciation. SVs can also cause dramatic changes in the regulation of multiple genes by disrupting chromatin domains and exposing certain promoters to certain enhancers for the first time. It was proposed that translocations and inversions perturbed TADs and thus created differences in promoter-enhancer connections between humans and mice that are responsible for differential regulation of genes involved in immune response between the two species (Gilbertson et al., 2022). SVs are otherwise strongly depleted from TAD boundaries and active chromatin states, suggesting that they are under negative selection (Fudenberg and Pollard, 2019). Single SVs often impact the expression of multiple genes, two on average in humans (Scott et al., 2021), suggesting that they frequently exert a pleiotropic effect on phenotypic diversity. Evidence of an SV with a strong and immediate effect on phenotype came from a recent experimental evolution study on nematode. Zhao et al. (2020) studied the genetic basis of adaptation to food sources in *Caenorhabditis elegans* and found a recombinant inbred line with increased fitness. They detected a complex SV as its genetic basis; this complex rearrangement caused duplication of a gene involved in exploration behavior and modified its expression. It was proposed that the SV occurred as a single genomic instability event and became fixed in a population because it provided a fitness advantage in a new environment. These findings highlight the potential of SVs in causing dramatic structural changes in the genome which can substantially and instantaneously affect phenotypes. The majority of such large events are expected to be deleterious. However, under specific circumstances, some variants may provide a strong selective advantage which would enable them to quickly rise in frequency within the population and even become fixed over a short evolutionary time.

Population-scale studies of SVs are strongly biased toward humans

Studying genetic variation in natural populations is crucial for understanding how genomes evolve. Assessing the degree of structural variation in various species and populations contributes to our general understanding of its role in evolutionary processes. Nevertheless, population-scale studies of SVs are heavily biased toward humans (Figure 1) and insights gained mainly from studies on human populations guide our general perception of structural variation (Box 1). However, modern humans have a specific population history that involved at least one severe bottleneck followed by rapid expansion and repeated founder effect (Watkins et al., 2001; Amos and Hoffman, 2010), which resulted in substantially lower

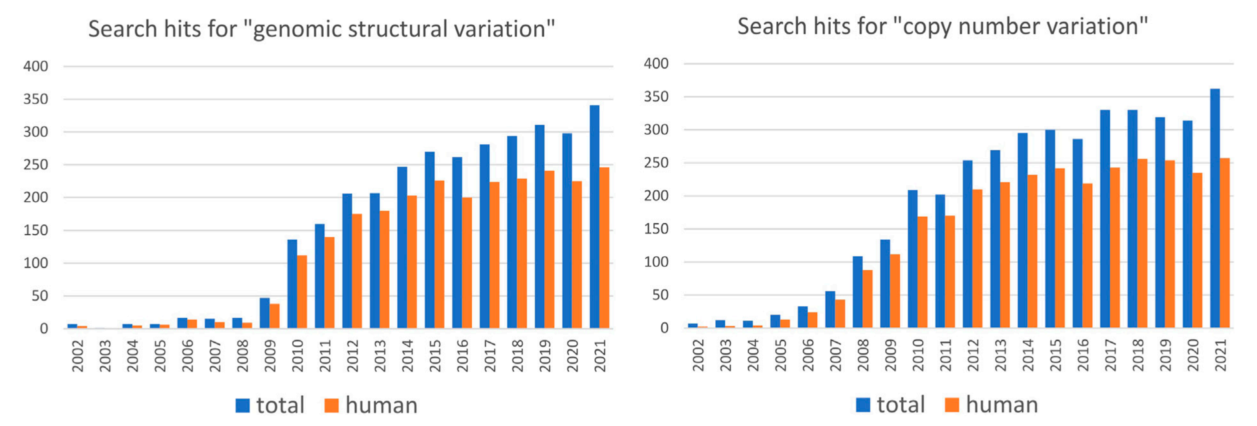


FIGURE 1
Number of articles per year containing indicated search term in PubMed in the period from 2002 to 2021. Search is performed for the combination of indicated term AND "population" (data in blue) or for indicated term AND "population" AND "human" (data in red). The data suggest that the majority of population-scale SV and CNV studies are based on humans.

genetic diversity compared to many other species. Moreover, genetic boundaries between human populations are often blurry, reflecting the frequent population movement and admixture. Humans are also characterized by a small effective population size (Tenesa et al., 2007; Park 2011), which is known to reduce the efficacy of natural selection and increase the influence of genetic drift. Thus, human populations by no means embody a "typical" evolutionary trajectory and more studies of SVs in non-human populations are needed to disentangle the roles that SVs play in evolution and ecological specialization. Based on growing evidence, SVs may be the key players of rapid adaptation to changing environments and naturally occurring examples of parallel evolution represent excellent opportunities to study the genetic architecture of rapid adaptation, such as adaptation to freshwater discussed above. The independence of studied populations that converged adaptive traits is desirable: the stronger the evidence that they independently evolved similar traits under the same selective pressure, the stronger the association with the underlying variant. Population studies in non-human species may provide more instances of such independent, parallel evolution as a framework for studying the role of SVs in adaptation and speciation. For instance, adaptation to the subterranean environment has been documented for many taxa, yet the impact of structural variation in this context is still unexplored. Numerous examples of parallel evolution can also be found in domesticated species, and evidence of SVs playing a part in trait evolution during domestication in plants and animals is emerging. For example, white coat color was independently selected for in sheep and goats - in both species, this trait is associated with duplication of the agouti signaling protein (*ASIP*) gene (Norris and Whan, 2008; Fontanesi et al., 2009). In plants, the loss of seed shattering was repeatedly selected for during domestication and is often associated with a deletion in gene *Sh1* in different cereal species (Lin et al., 2012; Choi et al., 2019). From an evolutionary point of view, domestication is a very specific process that usually involves a population bottleneck that substantially decreases genetic diversity and increases the frequency of domestication alleles (Gaut et al., 2018). It has been proposed that, at least in plants, deletions underly some of the crucial domestication traits, whereas during later stages of domestication (*i.e.* during diversification) various SV types facilitate local adaptation (Gaut et al., 2018; Lye and Purugganan, 2019). Hence, although they may provide some interesting examples of parallel evolution, domesticated species may not represent a general model for studying the role of SVs in evolution.

Challenges in the detection of structural variants

There is no doubt that sequencing technology based on short reads has tremendously advanced our knowledge of the prevalence of structural variation in populations and its impact on health and evolution over the past two decades. Numerous algorithms and approaches have been designed and employed to detect structural variants from short-read sequencing (SRS) data (reviewed in Kosugi et al., 2019). While many of them represent an improvement in some specific aspect, they all suffer from three basic problems, associated with technical limitations inherent to short reads. First, no single algorithm can detect SVs of all types and sizes. As shown by an exhaustive study that compared the performance of 69 existing algorithms for SV detection from WGS data, most algorithms perform best for particular SV types and, in some cases, for particular size ranges (Kosugi et al., 2019). Even when the same approach and the same algorithm is used, substantial differences may exist between samples that are not due to biological differences. For example, in the read-depth approach, lower coverage will lead to fewer SVs being identified, the power to detect smaller events will be compromised and neighboring SVs may collapse into single calls due to diminished resolution (Pezer et al., 2015). These problems make comparisons between studies difficult, as each approach applied to the same biological sample will result in a different set of SVs.

Second, the true positive rate of SRS-based methods is generally low while the false positive rates can be as high as 90%; again, both are heavily dependent on the size and type of SVs (Mahmoud et al., 2019). Differences in the processing of samples and data before SV calling can also strongly affect the accuracy of the final call set. For example, Khayat et al. (2021) found that sequencing centers and especially read mapping methods contribute significantly to variability between call sets. In particular, their results suggest that one-fifth of all calls represent false positives that are solely contributed by the mapper. These problems have major consequences on reproducibility and can greatly affect the interpretation between studies.

Third, methods based on SRS are unable to (accurately and reliably) identify SVs in repetitive genomic regions, stemming from the uncertainty of the true origin of reads that can be equally well mapped to multiple genomic positions. As a consequence, these problematic, repetitive regions are often omitted in genomic analyses. However, recent analyses based on long-read sequencing (LRS) technologies suggest that these regions may be the greatest source of variation. In human genomes, up to 90% of SVs (mostly smaller than 1 kbp) detected from LRS data were unknown from previous SRS-based analyses (Chaisson et al., 2015; Huddleston et al., 2017; Audano et al., 2019; Ebert et al., 2021; Quan et al., 2021). This means that SRS-based methods are blind to the vast majority of variation. This problem is particularly relevant in analyses of genomes with high repetitive DNA content such as in many plant species.

Despite its power to detect variation that is inaccessible to SRS, LRS has several drawbacks which directly limit its use in large population studies: it is more expensive, requires more input DNA, and has lower sample throughput than SRS (Ho et al., 2020). Consequently, not many population genomic studies based on long reads have emerged so far (Audano et al., 2019; Weissensteiner et al., 2020; Beyter et al., 2021; Quan et al., 2021; Yan et al., 2021; Rech et al., 2022). Majority of these studies employ a hybrid strategy which involves sequencing a smaller number of genomes by using long reads while the remaining samples are sequenced with short-read technology (Ho et al., 2020; De Coster et al., 2021; Quan et al., 2022). Structural variants identified by LRS in representative genomes can then be genotyped from SRS data in all other samples. This approach combines the advantages of both read-sequencing technologies: the power of LRS to discover multiple types and a wider size range of SVs (Quan et al., 2022), and the generally high genotyping precision of SRS-based algorithms (Kosugi et al., 2019). Even so, not all SVs that are detected from long reads can be accurately genotyped from short-read data, and as much as half remain invisible to it (Huddleston et al., 2017; Chakraborty et al., 2019; Ebert et al., 2021). Furthermore, LRS produces reads that are still insufficiently long to resolve all SVs. For instance, detection algorithms based on long reads that consider information on soft clipped reads and intra-read discordance are much worse at discovering CNVs larger than >100 kbp than are algorithms based on the read-depth approach from short reads (Kosugi et al., 2019). Hence, despite the advances made related to improved identification of smaller SVs by long reads, much of the most complex genomic regions remains inaccessible. Optical mapping is a technology of choice for resolving such regions as it generates molecules that can be over 1 Mb long and can therefore bridge larger repetitive regions (Ho et al., 2020). It has been successfully applied in some population studies which resolved previously undetected large SVs and identified novel genome content not found in the reference genome sequence (Levy-Sakin et al., 2019; Weissensteiner et al., 2020). However, optical mapping has several weak points, such as a high error rate, a lack of information on the actual sequence underlying the molecules, and the inherent inability to determine precise SV breakpoints. The widespread use of optical mapping is further hindered by lower throughput and the lack of alternative and publicly available tools for SV detection (see Li et al., 2017; Raeisi Dehkordi et al., 2021). Another promising technology that has the potential to detect large SVs and those in repetitive regions is high-throughput chromosome conformation capture (Hi-C). Hi-C is typically used for studying 3D genome interactions, and although several tools have been developed for SV discovery from Hi-C data, these are specifically designed for human genomes and are limited to the detection of SVs larger than 1 Mbp. Most recently, a framework named EagleC was developed that has the power to detect events down to 1 kb in any species genome, providing sufficient coverage (Wang et al., 2022b). This tool illustrates the potential of Hi-C application in SV discovery from large sample sets, and further developments in this direction will enable widespread and more comprehensive population-scale studies of SVs by use of Hi-C technology.

Towards more representative reference genome

In population studies, structural variants are most commonly detected from sequencing data by aligning reads to the reference genome sequence and identifying patterns of discordance in alignment. If the reference genome is contiguous, an average

Box 1 SVs in numbers

- 92,934 common structural variant regions in human populations; according to the NCBI Curated Common Structural Variants dataset (dbVar study accession nstd186; Lappalainen et al., 2013)
- 27,662 SVs detected per person, including STRs and other highly repetitive elements (Chaisson et al., 2019)
- 16 Mbp—The average amount of structural variation per person (Ebert et al., 2021)
- 3–15X—More base pairs are affected by SVs than by SNVs (Pang et al., 2010; Huddleston et al., 2017; Hämälä et al., 2021)
- 3–10X—Higher inter-individual genomic difference at SVs than at SNVs (Pang et al., 2010; Sudmant et al., 2015b)
- 4.8%–9.5% of the human genome is affected by CNVs (Zarrei et al., 2015)
- 0.29—Number of *de novo* SVs per generation (in regions of the genome accessible to short-read sequencing) or one new SV every two to eight live births (Collins et al., 2020; Belyeu et al., 2021)
- 6.8%—the largest estimated proportion of eQTLs caused by SVs (Chiang et al., 2017)

read depth of 10x is considered sufficient for population-scale comparisons (Collins et al., 2020). However, reference genomes assembled at the chromosome level are rarely available, which hampers studies in the majority of species. Even human genomes seem to contain large regions not present in the reference genome, as shown by studies based on optical mapping and long reads (Audano et al., 2019; Levy-Sakin et al., 2019; Ebert et al., 2021). They are not merely repetitive and non-functional, but also encompass genes and regulatory elements. These studies question the completeness and the representativeness of the human reference genome. The latest version of the human reference assembly, T2T-CHM13, succeeded in closing all gaps found in the previous GRCh38 assembly and indeed represents the first completely sequenced genome (Nurk et al., 2022). However, similar to the GRCh38, in which the majority of sequence originates from a single individual (Ballouz et al., 2019), T2T-CHM13 represents only one haplotype, and while it improves analysis of human genetic variation to some extent (Aganezov et al., 2022), it cannot fully capture the genetic diversity among populations. Approaches to remove reference bias have started to emerge, to improve accuracy in population-scale SV analyses. In 2019, Sherman et al. (2019), sequenced 910 individuals of African descent and used all unaligned reads to assemble contigs *de novo*. These collectively constituted 300 million base pairs of sequences that were missing from the reference genome and illustrated that a single reference genome is suboptimal for population-based studies. Instead, the creation of a comprehensive pan-genome was proposed, based on all distinct human populations that would much better capture all the DNA present in humans. In 2019, the Human Pangenome Project was initiated, funded by the US National Human Genome Research Institute (NHGRI), with a goal to provide a more accurate and diverse representation of global genomic variation through the creation of a more sophisticated human reference genome (Wang. et al., 2022a; Khamsi, 2022).

Pangenomes are superior to single reference genomes because they combine genomes from multiple individuals and thus better incorporate genomic polymorphism within a population, and they are becoming increasingly used for SV studies in humans and other species (Beyter et al., 2021; Ebert et al., 2021; Qin et al., 2021; Yan et al., 2021; Zhou et al., 2022; for a list of studies based on plant pangenomes see Yuan et al., 2021). Instead of being represented as a linear sequence, pangenomes are constructed as graphs to which sequencing reads are aligned (De Coster et al., 2021; Quan et al., 2022), enabling reliable genotyping of SVs by short reads in thousands of samples, which facilitates large population studies. However, approaches for graph-based genotyping are in their infancy, and tools for more efficient construction of complex graphs and alignment of reads to graphs are still under development (Quan et al., 2022).

The power of haplotype-resolved genomes

One of the major obstacles to a deeper understanding of SVs is the inability to accurately determine discrete SV alleles as it hinders evolutionary and population genetic studies of SVs, including analyses of allele frequency, estimations of the rates of recurrent mutation and incorporation of SVs in genome-wide association studies (Ebert et al., 2021; Saitou et al., 2022). This limitation can be overcome by resolving haplotypes. Studies that analyze haplotype-resolved genomes readily identify a substantial number of previously undetected SVs and additional genomic content not present in the reference genome (Huddleston et al., 2017; Wong et al., 2018; Chaisson et al., 2019; Levy-Sakin et al., 2019; Almarri et al., 2020; Ebert et al., 2021; Hämälä et al., 2021). Huddleston et al. (2017) sequenced genomes of two hydatiform moles - which are haploid; when they merged the two haploid genomes *in silico* to create an artificial diploid genome, over half of the heterozygous SVs were no longer detected from long-read sequencing data. This showed that the majority of SVs are not detectable unless the haplotype structure of the genomes is known and illustrates the importance of haploid resolution for the sensitivity of SV detection. However, determining the physical haplotype structure of genomes is yet not widely affordable and the haplotype-phasing methods are still immature, preventing their wider application in population-scale studies.

Variant interpretation

Over the past two decades, a sheer abundance of studies has demonstrated that structural variants are by far the most dominant form of genetic variation. While our ability to detect SVs has increased tremendously, for the largest part we are still unable to explain the functional consequences (Ho et al., 2020; Yan et al., 2021). For example, in the majority of population-scale studies, evidence on the adaptive role of SVs is inferred from associations between SV frequency and environmental/behavioral traits, however, rare studies provide evidence based on phenotypic assays, such as gene expression and protein level, or fitness (Perry et al., 2007; Maron et al., 2013; Ishikawa et al., 2019; Zhao et al., 2020). Experimental evolution provides a powerful means to study adaptation, yet it is limited to species with short generation times such as single-cell organisms. In more complex, multicellular organisms, it was proposed that integrating SVs with layered biological data is crucial for a more complete understanding of the impact of SVs (Ho et al., 2020). These may include but are not limited to analyses of transcriptome, epigenome, proteome, and 3D chromatin structure.

Concluding remarks

Recent years have shown that genome plasticity is even larger than it was anticipated more than 10 years ago. Long-read sequencing technologies have enabled the discovery of a wealth of structural variation in previously inaccessible genomic regions and continuous efforts provide increasing evidence that SVs play important roles in population divergence, local adaptation, and speciation. However, there is currently no approach that would allow simultaneous detection of all SVs, and even methods based on long reads fail in complex genomic regions such as long tandemly repetitive sequences and segmental duplications. Therefore, systemic assessments of SVs' contribution to evolution are primarily hindered by the high cost of analyzing a large number of individuals to enable population-scale studies, and by the necessity to employ multiple available technologies, in order to capture all types of SVs and achieve greater resolution of SV detection. Without a such comprehensive approach, the investigations are limited to SVs of a particular size range or types. Pangenome assemblies provide a route to avoid costly sequencing by long reads and enable genotyping from short reads mapped on a reference genome that is derived from several individuals, representative of multiple populations. Investment in efforts to construct pangenomes in a multitude of species will enable more reliable and comprehensive SV detection and genotyping on a larger scale. The number of detected SVs is expected to increase further with the improvement of haplotype-phasing methods, and the wider application of such methods is expected to greatly advance our understanding of the impact of SVs on evolution.

Author contributions

ZP conceptualized the manuscript. IP and ZP wrote the manuscript.

Funding

This work was supported by the Croatian Science Foundation under Grant UIP-2019-04-7898.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., et al. (2022). A complete reference genome improves analysis of human genetic variation. *Science* 376, eabl3533. doi:10.1126/science.abl3533

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi:10.1038/nrg2958

Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., et al. (2020). Population structure, stratification, and introgression of human structural variation. *Cell* 182, 189–199. doi:10.1016/j.cell.2020.05.024

Amos, W., and Hoffman, J. I. (2010). Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc. Biol. Sci.* 277, 131–137. doi:10.1098/rspb.2009.1473

Aplan, P. D. (2006). Causes of oncogenic chromosomal translocation. *Trends Genet.* 22, 46–55. doi:10.1016/j.tig.2005.10.002

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell* 176, 663–675. e19. doi:10.1016/j.cell.2018.12.019

Baek, Y., Bobadilla, L. K., Giacomini, D. A., Montgomery, J. S., Murphy, B. P., and Tranel, P. J. (2021). Evolution of glyphosate-resistant weeds. *Rev. Environ. Contam. Toxicol.* 255, 93–128. doi:10.1007/398_2020_55

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20, 159. doi:10.1186/s13059-019-1774-4

Belyeu, J. R., Brand, H., Wang, H., Zhao, X., Pedersen, B. S., Feusier, J., et al. (2021). De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2, 396 families. *Am. J. Hum. Genet.* 108, 597–607. doi:10.1016/j.ajhg.2021.02.012

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., et al. (2021). Long-read sequencing of 3, 622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* 53, 779–786. doi:10.1038/s41588-021-00865-4

Boettger, L. M., Salem, R. M., Handsaker, R. E., Peloso, G. M., Kathiresan, S., Hirschhorn, J. N., et al. (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* 48, 359–366. doi:10.1038/ng.3510

Carvalho, C. M., and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* 17, 224–238. doi:10.1038/nrg.2015.25

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611. doi:10.1038/nature13907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10, 1784. doi:10.1038/s41467-018-08148-z

Chakraborty, M., Emerson, J. J., Macdonald, S. J., and Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* 10, 4872. doi:10.1038/s41467-019-12884-1

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., et al. (2017). The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699. doi:10.1038/ng.3834

Choi, J. Y., Zaidem, M., Gutaker, R., Dorph, K., Singh, R. K., and Purugganan, M. D. (2019). The complex geography of domestication of the African rice Oryza glaberrima. *PLoS Genet.* 15, e1007414. doi:10.1371/journal.pgen.1007414

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., et al. (2020). A structural variation reference for medical and population genetics. *Nature* 581, 444–451. doi:10.1038/s41586-020-2287-8

De Coster, W., Weissensteiner, M. H., and Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nat. Rev. Genet.* 22, 572–587. doi:10.1038/s41576-021-00367-3

Díaz, A., Zikhali, M., Turner, A. S., Isaac, P., and Laurie, D. A. (2012). Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One* 7, e33234. doi:10.1371/journal.pone.0033234

Dopman, E. B., and Hartl, D. L. (2007). A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19920–19925. doi:10.1073/pnas.0709888104

Dumas, L., Kim, Y. H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J. R., et al. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* 17, 1266–1277. doi:10.1101/gr.6557307

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372, eabf7117. doi:10.1126/science.abf7117

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97. doi:10.1038/nrg1767

Fontanesi, L., Beretti, F., Riggio, V., GómezGonzález, E., Dall'Olio, S., Davoli, R., et al. (2009). Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet. Genome Res.* 126, 333–347. doi:10.1159/000268089

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* 2, E207. doi:10.1371/journal.pbio.0020207

Fudenberg, G., and Pollard, K. S. (2019). Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2175–2180. doi:10.1073/pnas.1808631116

Gaines, T. A., Zhang, W., Wang, D., Bukun, B., Chisholm, S. T., Shaner, D. L., et al. (2010). Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proc. Natl. Acad. Sci. U. S. A.* 107, 1029–1034. doi:10.1073/pnas.0906649107

Garagna, S., Page, J., Fernandez-Donoso, R., Zuccotti, M., and Searle, J. B. (2014). The robertsonian phenomenon in the house mouse: Mutation, meiosis and speciation. *Chromosoma* 123, 529–544. doi:10.1007/s00412-014-0477-6

Garg, P., Jadhav, B., Lee, W., Rodriguez, O. L., Martin-Trujillo, A., and Sharp, A. J. (2022). A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. *Am. J. Hum. Genet.* 109, 1065–1076. doi:10.1016/j.ajhg.2022.04.016

Garg, P., Martin-Trujillo, A., Rodriguez, O. L., Gies, S. J., Hadelia, E., Jadhav, B., et al. (2021). Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am. J. Hum. Genet.* 108, 809–824. doi:10.1016/j.ajhg.2021.03.016

Gaut, B. S., Seymour, D. K., Liu, Q., and Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. *Nat. Plants* 4, 512–520. doi:10.1038/s41477-018-0210-1

Gilbertson, S. E., Walter, H. C., Gardner, K., Wren, S. N., Vahedi, G., and Weinmann, A. S. (2022). Topologically associating domains are disrupted by evolutionary genome rearrangements forming species-specific enhancer connections in mice and humans. *Cell Rep.* 39, 110769. doi:10.1016/j.celrep.2022.110769

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., de Pamphilis, C. W., and Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2102914118. doi:10.1073/pnas.2102914118

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., et al. (2015). Large multiallelic copy number variations in humans. *Nat. Genet.* 47, 296–303. doi:10.1038/ng.3200

Harringmeyer, O. S., and Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat. Ecol. Evol.* doi:10.1038/s41559-022-01890-0

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 10, 551–564. doi:10.1038/nrg2593

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., et al. (2009). Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.* 41, 424–429. doi:10.1038/ng.345

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. *Nat. Rev. Genet.* 21, 171–189. doi:10.1038/s41576-019-0180-9

Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., et al. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* 366, eaax2083. doi:10.1126/science.aax2083

Huang, Y., Feulner, P. G. D., Eizaguirre, C., Lenz, T. L., Bornberg-Bauer, E., Milinski, M., et al. (2019). Genome-wide genotype-expression relationships reveal both copy number and single nucleotide differentiation contribute to differential gene expression between stickleback ecotypes. *Genome Biol. Evol.* 11, 2344–2359. doi:10.1093/gbe/evz148

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685. doi:10.1101/gr.214007.116

Ishikawa, A., Kabeya, N., Ikeya, K., Kakioka, R., Cech, J. N., Osada, N., et al. (2019). A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science* 364, 886–889. doi:10.1126/science.aau5656

Ishikawa, A., Yamanouchi, S., Iwasaki, W., and Kitano, J. (2022). Convergent copy number increase of genes associated with freshwater colonization in fishes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 377, 20200509. doi:10.1098/rstb.2020.0509

Iskow, R. C., Gokcumen, O., and Lee, C. (2012). Exploring the role of copy number variants in human adaptation. *Trends Genet.* 28, 245–257. doi:10.1016/j.tig.2012.03.002

Keel, B. N., Lindholm-Perry, A. K., and Snelling, W. M. (2016). Evolutionary and functional features of copy number variation in the cattle genome. *Front. Genet.* 7, 207. doi:10.3389/fgene.2016.00207

Khamsi, R. (2022). A more-inclusive genome project aims to capture all of human diversity. *Nature* 603, 378–381. doi:10.1038/d41586-022-00726-y

Khayat, M. M., Sahraeian, S. M. E., Zarate, S., Carroll, A., Hong, H., Pan, B., et al. (2021). Hidden biases in germline structural variant detection. *Genome Biol.* 22, 347. doi:10.1186/s13059-021-02558-x

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488, 471–475. doi:10.1038/nature11396

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20, 117. doi:10.1186/s13059-019-1720-5

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., et al. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res.* 41, D936–D941. doi:10.1093/nar/gks1213

Lauer, S., and Gresham, D. (2019). An evolving view of copy number variants. *Curr. Genet.* 65, 1287–1295. doi:10.1007/s00294-019-00980-0

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., et al. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* 10, 1025. doi:10.1038/s41467-019-08992-7

Li, L., Leung, A. K., Kwok, T. P., Lai, Y. Y. Y., Pang, I. K., Chung, G. T., et al. (2017). OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biol.* 18, 230. doi:10.1186/s13059-017-1356-2

Lin, Z., Li, X., Shannon, L. M., Yeh, C. T., Wang, M. L., Bai, G., et al. (2012). Parallel domestication of the Shattering1 genes in cereals. *Nat. Genet.* 44, 720–724. doi:10.1038/ng.2281

Lye, Z. N., and Purugganan, M. D. (2019). Copy number variation in domestication. *Trends Plant Sci.* 24, 352–365. doi:10.1016/j.tplants.2019.01.003

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biol.* 20, 246. doi:10.1186/s13059-019-1828-7

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5241–5246. doi:10.1073/pnas.1220766110

Martin, G., Baurens, F. C., Hervouet, C., Salmon, F., Delos, J. M., Labadie, K., et al. (2020). Chromosome reciprocal translocations have accompanied subspecies evolution in bananas. *Plant J.* 104, 1698–1711. doi:10.1111/tpj.15031

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* 7, 233–245. doi:10.1038/nrc2091

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

North, H. L., Caminade, P., Severac, D., Belkhir, K., and Smadja, C. M. (2020). The role of copy-number variation in the reinforcement of sexual isolation between the two European subspecies of the house mouse. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375, 20190540. doi:10.1098/rstb.2019.0540

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. doi:10.1126/science.abj6987

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11, R52. doi:10.1186/gb-2010-11-5-r52

Park, L. (2011). Effective population size of current human population. *Genet. Res.* 93, 105–114. doi:10.1017/S0016672310000558

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39, 1256–1260. doi:10.1038/ng2123

Pezer, Ž., Harr, B., Teschke, M., Babiker, H., and Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus* domesticus) reveal three conserved genes with major population-specific expansions. *Genome Res.* 25, 1114–1124. doi:10.1101/gr.187187.114

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., et al. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* 44, 548–559. doi:10.1016/j.bj.2021.02.003

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184, 3542–3558.e16. doi:10.1016/j.cell.2021.04.046

Quan, C., Li, Y., Liu, X., Wang, Y., Ping, J., Lu, Y., et al. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biol.* 22, 159. doi:10.1186/s13059-021-02382-3

Quan, C., Lu, H., Lu, Y., and Zhou, G. (2022). Population-scale genotyping of structural variation in the era of long-read sequencing. *Comput. Struct. Biotechnol. J.* 20, 2639–2647. doi:10.1016/j.csbj.2022.05.047

Radke, D. W., and Lee, C. (2015). Adaptive potential of genomic structural variation in human and mammalian evolution. *Brief. Funct. Genomics* 14, 358–368. doi:10.1093/bfgp/elv019

Raeisi Dehkordi, S., Luebeck, J., and Bafna, V. (2021). FaNDOM: Fast nested distance-based seeding of optical maps. *Patterns* 2, 100248. doi:10.1016/j.patter.2021.100248

Rech, G. E., Radío, S., Guirao-Rico, S., Aguilera, L., Horvath, V., Green, L., et al. (2022). Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. *Nat. Commun.* 13, 1948. doi:10.1038/s41467-022-29518-8

Rigau, M., Juan, D., Valencia, A., and Rico, D. (2019). Intronic CNVs and gene expression variation in human populations. *PLoS Genet.* 15, e1007902. doi:10.1371/journal.pgen.1007902

Rinker, D. C., Specian, N. K., Zhao, S., and Gibbons, J. G. (2019). Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. *Proc. Natl. Acad. Sci. U. S. A.* 116, 13446–13451. doi:10.1073/pnas.1901093116

Saitou, M., Masuda, N., and Gokcumen, O. (2022). Similarity-based analysis of allele frequency distribution among multiple populations identifies adaptive genomic structural variants. *Mol. Biol. Evol.* 39, msab313. doi:10.1093/molbev/msab313

Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Brief. Funct. Genomics* 13, 296–307. doi:10.1093/bfgp/elu016

Schrider, D. R., Navarro, F. C., Galante, P. A., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., et al. (2013). Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* 9, e1003242. doi:10.1371/journal.pgen.1003242

Scott, A. J., Chiang, C., and Hall, I. M. (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.* 31, 2249–2257. doi:10.1101/gr.275488.121

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 51, 30–35. doi:10.1038/s41588-018-0273-y

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants* 6, 34–45. doi:10.1038/s41477-019-0577-7

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nat. Rev. Genet.* 19, 453–467. doi:10.1038/s41576-018-0007-0

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* 23, 1373–1382. doi:10.1101/gr.158543.113

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., et al. (2015a). Global diversity, population stratification, and selection of human copy-number variation. *Science*. 349. aab3761. doi:10.1126/science.aab3761

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015b). An integrated map of structural variation in 2, 504 human genomes. *Nature* 526, 75–81. doi:10.1038/nature15394

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, 520–526. doi:10.1101/gr.6023607

Thomas, G. W. C., Wang, R. J., Nguyen, J., AlanHarris, R., Raveendran, M., Rogers, J., et al. (2021). Origins and long-term patterns of copy-number variation in rhesus macaques. *Mol. Biol. Evol.* 38, 1460–1471. doi:10.1093/molbev/msaa303

Vickrey, A. I., Bruders, R., Kronenberg, Z., Mackey, E., Bohlender, R. J., Maclary, E. T., et al. (2018). Introgression of regulatory alleles and a missense coding mutation drive plumage pattern diversity in the rock pigeon. *Elife* 7, e34803. doi:10.7554/eLife.34803

Wang, R. J., Thomas, G. W. C., Raveendran, M., Harris, R. A., Doddapaneni, H., Muzny, D. M., et al. (2020). Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. *Genome Res.* 30, 826–834. doi:10.1101/gr.255174.119

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., et al. (2022a). The human pangenome project: A global resource to map genomic diversity. *Nature* 604, 437–446. doi:10.1038/s41586-022-04601-8

Wang, X., Luan, Y., and Yue, F. (2022b). EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci. Adv.* 8, eabn9215. doi:10.1126/sciadv.abn9215

Watkins, W. S., Ricker, C. E., Bamshad, M. J., Carroll, M. L., Nguyen, S. V., Batzer, M. A., et al. (2001). Patterns of ancestral human diversity: An analysis of alu-insertion and restriction-site polymorphisms. *Am. J. Hum. Genet.* 68, 738–752. doi:10.1086/318793

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K. J., Knief, U., Heim, W., et al. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nat. Commun.* 11, 3403. doi:10.1038/s41467-020-17195-4

Wellenreuther, M., and Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* 33, 427–440. doi:10.1016/j.tree.2018.04.002

Wong, K. H. Y., Levy-Sakin, M., and Kwok, P. Y. (2018). De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun.* 9, 3040. doi:10.1038/s41467-018-05513-w

Würschum, T., Boeven, P. H., Langer, S. M., Longin, C. F., and Leiser, W. L. (2015). Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genet.* 16, 96. doi:10.1186/s12863-015-0258-0

Xia, Y., Yuan, X., Luo, W., Yuan, S., and Zeng, X. (2020). The origin and evolution of chromosomal reciprocal translocation in quasipaa boulengeri (Anura, dicroglossidae). *Front. Genet.* 10, 1364. doi:10.3389/fgene.2019.01364

Xu, L., Hou, Y., Bickhart, D. M., Zhou, Y., Hayel, H. A., Song, J., et al. (2016). Population-genetic properties of differentiated copy number variations in cattle. *Sci. Rep.* 6, 23161. doi:10.1038/srep23161

Yan, S. M., Sherman, R. M., Taylor, D. J., Nair, D. R., Bortvin, A. N., Schatz, M. C., et al. (2021). Local adaptation and archaic introgression shape global diversity at human structural variant loci. *Elife* 10, e67615. doi:10.7554/eLife.67615

Yuan, Y., Bayer, P. E., Batley, J., and Edwards, D. (2021). Current status of structural variation studies in plants. *Plant Biotechnol. J.* 19, 2153–2163. doi:10.1111/pbi.13646

Yuste-Lisbona, F. J., Fernández-Lozano, A., Pineda, B., Bretones, S., Ortíz-Atienza, A., García-Sogo, B., et al. (2020). ENO regulates tomato fruit size through the floral meristem development network. *Proc. Natl. Acad. Sci. U. S. A.* 117, 8187–8195. doi:10.1073/pnas.1913688117

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nat. Rev. Genet.* 16, 172–183. doi:10.1038/nrg3871

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi:10.1146/annurev.genom.9.081307.164217

Zhang, L., Reifová, R., Halenková, Z., and Gompert, Z. (2021). How important are structural variants for speciation? *Genes* 12, 1084. doi:10.3390/genes12071084

Zhang, W., and Tautz, D. (2022). Tracing the origin and evolutionary fate of recent gene retrocopies in natural populations of the house mouse. *Mol. Biol. Evol.* 39, msab360. doi:10.1093/molbev/msab360

Zhao, Y., Long, L., Wan, J., Biliya, S., Brady, S. C., Lee, D., et al. (2020). A spontaneous complex structural variant in rcan-1 increases exploratory behavior and laboratory fitness of *Caenorhabditis elegans*. *PLoS Genet.* 16, e1008606. doi:10.1371/journal.pgen.1008606

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606, 527–534. doi:10.1038/s41586-022-04808-9

Zong, S. B., Li, Y. L., and Liu, J. X. (2021). Genomic architecture of rapid parallel adaptation to fresh water in a wild fish. *Mol. Biol. Evol.* 38, 1317–1329. doi:10.1093/molbev/msaa290

Zuellig, M. P., and Sweigart, A. L. (2018). Gene duplicates cause hybrid lethality between sympatric species of Mimulus. *PLoS Genet.* 14, e1007130. doi:10.1371/journal.pgen.1007130

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.993959

OPEN ACCESS

EDITED BY
Stéphane Joost,
Swiss Federal Institute of Technology Lausanne, Switzerland

REVIEWED BY
Mehmet Gültas,
South Westphalia University of Applied Sciences, Germany
Hugo Oswaldo Toledo-Alvarado,
National Autonomous University of Mexico, Mexico
Mario Barbato,
Catholic University of the Sacred Heart, Italy

*CORRESPONDENCE
Gudrun A. Brockmann,
✉ gudrun.brockmann@hu-berlin.de

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

RECEIVED 14 July 2022
ACCEPTED 28 November 2022
PUBLISHED 04 January 2023

CITATION
Neumann GB, Korkuć P, Arends D, Wolf MJ, May K, König S and Brockmann GA (2023), Genomic diversity and relationship analyses of endangered German Black Pied cattle (DSN) to 68 other taurine breeds based on whole-genome sequencing.
Front. Genet. 13:993959.
doi: 10.3389/fgene.2022.993959

Genomic diversity and relationship analyses of endangered German Black Pied cattle (DSN) to 68 other taurine breeds based on whole-genome sequencing

Guilherme B. Neumann[1], Paula Korkuć[1], Danny Arends[1,2], Manuel J. Wolf[3], Katharina May[3], Sven König[3] and Gudrun A. Brockmann[1]*

[1]Animal Breeding Biology and Molecular Genetics, Albrecht Daniel Thaer-Institute for Agricultural and Horticultural Sciences, Humboldt-Universität zu Berlin, Berlin, Germany, [2]Department of Applied Sciences, Northumbria University, Newcastle Upon Tyne, United Kingdom, [3]Institute of Animal Breeding and Genetics, Justus-Liebig-Universität, Giessen, Germany

German Black Pied cattle (Deutsches Schwarzbuntes Niederungsrind, DSN) are an endangered dual-purpose cattle breed originating from the North Sea region. The population comprises about 2,500 cattle and is considered one of the ancestral populations of the modern Holstein breed. The current study aimed at defining the breeds closest related to DSN cattle, characterizing their genomic diversity and inbreeding. In addition, the detection of selection signatures between DSN and Holstein was a goal. Relationship analyses using fixation index (F_{ST}), phylogenetic, and admixture analyses were performed between DSN and 68 other breeds from the 1000 Bull Genomes Project. Nucleotide diversity, observed heterozygosity, and expected heterozygosity were calculated as metrics for genomic diversity. Inbreeding was measured as excess of homozygosity (F_{Hom}) and genomic inbreeding (F_{RoH}) through runs of homozygosity (RoHs). Region-wide F_{ST} and cross-population-extended haplotype homozygosity (XP-EHH) between DSN and Holstein were used to detect selection signatures between the two breeds, and RoH islands were used to detect selection signatures within DSN and Holstein. DSN showed a close genetic relationship with breeds from the Netherlands, Belgium, Northern Germany, and Scandinavia, such as Dutch Friesian Red, Dutch Improved Red, Belgian Red White Campine, Red White Dual Purpose, Modern Angler, Modern Danish Red, and Holstein. The nucleotide diversity in DSN (0.151%) was higher than in Holstein (0.147%) and other breeds, e.g., Norwegian Red (0.149%), Red White Dual Purpose (0.149%), Swedish Red (0.149%), Hereford (0.145%), Angus (0.143%), and Jersey (0.136%). The F_{Hom} and F_{RoH} values in DSN were among the lowest. Regions with high F_{ST} between DSN and Holstein, significant XP-EHH regions, and RoH islands detected in both breeds harbor candidate genes that were previously reported for milk, meat, fertility, production, and health traits, including one QTL detected in DSN for endoparasite infection resistance. The selection signatures between DSN and Holstein provide evidence of regions responsible for the dual-purpose properties of DSN and the milk type of Holstein. Despite the small population size, DSN has a high level of diversity and low inbreeding. F_{ST} supports its relatedness to breeds from the same geographic origin and provides information on potential gene pools that could be used to maintain diversity in DSN.

KEYWORDS

***Bos taurus*, Holstein, whole-genome sequencing, inbreeding, phylogenetic analysis, runs of homozygosity, selection signatures, 1000 Bull Genomes project**

1 Introduction

Autochthonous populations are a crucial source of genetic diversity for the conservation of livestock species harboring important local adaptations (Medugorac et al., 2011). However, local breeds are typically less productive than intensively selected high-performing breeds. Consequently, keeping local breeds is less profitable (Gandini et al., 2007; Hiemstra et al., 2010). Thus, many local populations have been replaced by more profitable breeds which dramatically reduced the herd size of local populations. According to the Food and Agriculture Organization of the United Nations (FAO), 84% of all local breeds in Europe were considered at risk of extinction in 2021 (FAO, 2021).

This trend can also be observed for the German Black Pied cattle population (“Deutsches Schwarzbuntes Niederungsrind”, DSN). DSN is an endangered dual-purpose cattle breed from Germany (GEH e.V., 2020). Its initial farming dates back to the 18th century in the North Sea region of Germany and the Netherlands, where black and white animals were kept (today named DSN in Germany, and Dutch Friesian in the Netherlands). From there, black and white cattle were exported to North America and other parts of Europe. Strong selection on milk yield and dairy character resulted in the high-yielding dairy breed named Holstein Friesian (Holstein). These high-yielding cattle were brought back to Europe in the mid-1960s and rapidly became one of the main dairy cattle breeds worldwide. As a boomerang, DSN cattle were replaced almost entirely by Holstein (Brade, 2011; Brade and Brade, 2013). In 2020, the number of DSN herdbook cows was 2,452 (TGRDEU, 2021). According to the Society for the Conservation of Old and Endangered Livestock Breeds (GEH), populations with an effective population size (Ne) below 200 should be kept as genetic resources (Stier, 2021). For DSN, Ne was estimated as 85 (Jaeger et al., 2018). For that reason, this population became a genetic reserve in 1972, as a resource for the future of livestock breeding in Germany. This decision is in agreement with the Global Plan of Action for Animal Genetic Resources (FAO, 2007), in which it is stated that local breeds represent genetic resources that contain important alleles for the adaptation to local conditions (Hoffmann, 2011; Medugorac et al., 2011; Boettcher et al., 2014; Biscarini et al., 2015).

In this context, the breeding goal in DSN is to conserve the typical and beneficial characteristics of this breed, which are high robustness, fertility, longevity, resistance to multiple diseases, calm temperament, correct positioning of feet and legs, as well as high roughage feed intake capacity, making it suitable for grazing (BRS, 2021). Those characteristics confirm an advantage of DSN to organic farming. Not all those traits, however, are yet fully described in DSN and genetic variants affecting those traits are unknown. So far, studies on milk production (Korkuć et al., 2021), mastitis resistance (Meier et al., 2020), endoparasites infection resistance (May et al., 2019), and fertility traits (Wolf et al., 2021) identified some candidate variants and genes affecting those traits. To improve the identification of DSN-typical DNA variants underlying the phenotypic variance, a customized SNP chip was designed for DSN (Neumann et al., 2021). This DSN-specific SNP chip is currently used to support the characterization of genomic diversity and the identification of association between genetic variants and diverse phenotypes.

For an effective conservation plan, genomic diversity measurements are necessary to evaluate and ensure a minimum pool of genetic variants that provides sufficient adaptation capacity to changing environments and prevents inbreeding depression (Kristensen et al., 2015). Besides, typical measures of heterozygosity (observed and expected) and excess, of homozygosity, genome-wide nucleotide diversity, for example, has been shown to be useful when evaluating the genetic diversity status of a given population (Kardos et al., 2021). The nucleotide diversity is a region-wide metric that is used to quantify the degree of polymorphisms within a population. Additionally, runs of homozygosity (RoH) (McQuillan et al., 2008) have been used to detect inbreeding and signatures of selection on a genome-wide level (Mészáros et al., 2015; Mastrangelo et al., 2016; Gorssen et al., 2021).

Besides the joint historic origin of DSN and Holstein population (Naderi et al., 2020), little is known about which of the breeds that are maintained today are the most closely related to DSN. Due to the small population size of DSN, the risk of increasing inbreeding and losing diversity is high. Taking this into account, the breeds identified to be the most closely related to DSN on the genome level could potentially serve as genomic sources for maintaining and improving the genomic diversity within DSN and *vice versa*. If we look at the history and customs of people, we expect that breeds from the same region near the North Sea show a high level of genetic proximity (Brade and Brade, 2013; Felius et al., 2014). For that reason, analyses of the fixation index (F_{ST}), phylogeny, and admixture could provide important information about the relationship between DSN and other populations from the same or different geographical locations.

Information on the genomic diversity of local breeds such as DSN and its relationship to other breeds can be taken into account in livestock breeding for maintaining the diversity within a breed and improving resilience. In the case of inbreeding depression or the spread of tropical diseases due to climate change, for example, known genomic regions improving disease resistance could be used for genetic rescue programs (Medugorac et al., 2011; Kristensen et al., 2015; Kardos et al., 2021). This would be especially useful for closely related breeds such as DSN and Holstein. Since DSN cattle are maintained under less selection intensity than Holstein, we expect that DSN cattle contain sequence variants that increase phenotypic plasticity and resilience.

The aim of the current work was to characterize the genomic diversity of DSN and define the breeds most closely related to DSN. To obtain a better understanding of the genomic diversity of DSN, we calculated and compared genomic measurements (observed heterozygosity, expected heterozygosity, excess of homozygosity, nucleotide diversity, and genomic inbreeding) within DSN and between breeds. For the analyses, we used whole-genome sequencing data of 302 DSN cattle together with the sequence variants of 68 other taurine breeds obtained from the 1000 Bull Genomes Project (Hayes and Daetwyler, 2019). In order to support the diversity analyses and the definition of the breeds most closely related to DSN, relationship analyses (F_{ST}, phylogeny, and admixture) were performed between DSN and the other breeds. To detect regions of high differentiation between DSN and Holstein, we searched for signatures of selection within and between these two breeds.

2 Materials and methods

2.1 Genomic data

Sequence variants from whole-genome sequencing data of 302 DSN animals (Neumann et al., 2021) and 1,388 animals of additional 68 taurine breeds and one Auroch (*Bos primigenius*) from the 1000 Bull Genomes Project (Run 9) (Hayes and Daetwyler, 2019) were used in this study. Among the sequenced DSN cattle, there were 12 key ancestors of the last 44 to 20 years. Data pre-processing, sequence read mapping, variant discovery and recalibration for DSN have been described previously (Neumann et al., 2021). Basically, we followed the same pipeline guidelines as for the data from the 1000 Bull Genomes Project, using the *Bos taurus* genome version ARS-UCD1.2 as reference (Rosen et al., 2020). From the 1000 Bull Genomes Project data, only breeds with at least five animals with a minimum average read depth of 8-fold were used (Table 1). The maximum number of animals per breed was restricted to 30, whereas these 30 animals were randomly selected. Exceptions were made for Holstein, where 150 animals were selected (with 30 animals randomly selected per country), and Red and White Dual Purpose, where all 42 available animals were used due to pre-knowledge about their genetic proximity to DSN (Neumann et al., 2021). The sequence variants from the 302 DSN and the other breeds were merged using BCFtools v.1.9 (Narasimhan et al., 2016). A total of 79,019,242 biallelic autosomal variants (72,329,983 SNPs and 6,689,259 indels) occurring in the tranche 99% (from the Variant Recalibration performed by 1000 Bull Genomes Project) and with a call rate ≥ 0.95, were considered in our analyses (Supplementary Figure S1).

For the phylogenetic tree and admixture analyses, the available 79,019,242 sequence variants were pruned (--indep-pairwise) with PLINK v2 (Purcell et al., 2007) using a r^2 threshold of 0.6, a window size of 50 SNPs, and a step-size of 5 SNPs to 23,059,286 variants. The same parameters for pruning were also used for other cattle analyses based on WGS (Xia et al., 2021; Zhang et al., 2022), except for a higher, more conservative r^2 of 0.6, used at this point to keep variants in medium LD. All 68 breeds were used in the phylogenetic tree and in the F_{ST} analyses. Subsequently, the 20 breeds most closely related to DSN according to the estimated F_{ST} values were used in the admixture analysis.

Diversity and inbreeding measures were calculated for 24 breeds (Table 1) which had at least 25 animals. According to the FAO recommendation, 25 is the minimum sample count for a precise genetic diversity description of a population (FAO, 2011). For these 24 breeds, the initial 79,019,242 sequence variants were filtered down to 34,856,428 variants segregating among the 1,118 animals of those 24 breeds.

2.2 Relationship analyses

2.2.1 Phylogenetic tree

To study the relationship between the available breeds and to detect the most closely related breeds to DSN, a genome-wide phylogenetic tree was built for all *Bos taurus* autosomes (BTA). Based on 23,059,286 pruned sequence variants, Manhattan distances between animals were calculated and the Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm implemented in the biotite library v0.35.0 (Kunzmann and Hamacher, 2018) in Python was used

TABLE 1 Number of animals selected per breed and analyses in which they were included. R stands for 'Relationship' including phylogeny and F_{ST} (all breeds), A stands for 'Admixture' (20 breeds with highest F_{ST}), and D for 'Diversity' (24 breeds with ≥ 25 animals).

Breed	n	Analyses	Breed	n	Analyses	Breed	n	Analyses
Abondance	9	R	Gelbvieh	30	R, D	Red and White Dual Purpose	42	R, A, D
Altai	20	R	German Red Angler	10	R, A	Ringamålako	8	R
Angus	30	R, D	Groningen White Headed	10	R	Romagnola	21	R
Angus Red	18	R	Guernsey	20	R	Rotes Höhenvieh	6	R
Aubrac	5	R	Hanwoo	20	R	Salers	18	R
Auroch	1	R	Hereford	30	R, D	Scottish Highland	7	R
Ayrshire Finnish	30	R, D	Holstein	150	R, A, D	Shorthorn	29	R, D
Belgian Red White Campine	10	R, A	Holstein Red	17	R, A	Simmental	30	R, D
Blonded Aquitaine	30	R, D	Jersey	30	R, D	Swedish Red	30	R, D
Brown Swiss	30	R, D	Kalmykian	10	R	Swedish Red Polled	6	R, A
Buryat	19	R	Kholmogory	30	R, A, D	Tarentaise	11	R
Buša	10	R	Limousin	30	R, D	Traditional Danish Red	9	R
Charolais	30	R, A, D	Maine Anjou	20	R	Tyrolean Grey	17	R
Chianina	15	R	Marchigiana	9	R	Ukrainian Grey	8	R
Deep Red Cattle	9	R, A	Menggu	10	R	Väneko	11	R
DSN	302	R, A, D	Modern Angler	20	R, A	Vorderwälder	13	R
Dutch Belted	11	R, A	Modern Danish Red	28	R, A, D	Wagyu	28	R, D
Dutch Friesian Red	11	R, A	Montbeliarde	30	R, D	Western Finncattle	15	R
Dutch Improved Red	9	R, A	Normande	30	R, D	West Vlaams Rood	11	R, A
Eastern Belgian Red White	7	R, A	Northern Finncattle	19	R	Yakut	30	R, D
Eastern Finncattle	15	R, A	Norwegian Red	29	R, A, D	Yanbian	10	R
Eastern Flanders White Red	12	R, A	Original Braunvieh	30	R, D	Yaroslavl	22	R
Fjäll	17	R	Podolian Serbia	10	R	Total 1,691		
Fleckvieh	30	R, D	Polish Red	7	R, A			

for clustering. The phylogenetic tree was visualized using iTOL v6 (Letunic and Bork, 2019). Individuals of the same breed were collapsed to a branch labelled with the breed's name. One auroch (*Bos primigenius*) was added as an outgroup in order to root the tree. Animals of a specific breed clustering outside the expected breed branch were removed. In the case of multiple clusters for a single breed, only the biggest cluster was kept. In addition, to detect migration events, a maximum likelihood tree allowing two migration events with bootstrap in blocks of 500 variants was built using TreeMix v1.13 (Pickrell and Pritchard, 2012). Number of migrations (m) was defined based on the Evanno method (Evanno et al., 2005) as implemented in the R package OptM v0.1.6 (Fitak, 2021). Different m from 2 to 6 were tested and a maximum Δm = 8.66 was estimated when m = 2 edges were selected.

2.2.2 F_{ST} calculation

Pairwise F_{ST} values between DSN and the other 68 breeds were calculated using variants segregating in either DSN or the other breed from the initial dataset of 79,019,242 sequence variants. F_{ST} values were estimated based on Hudson's method (Hudson et al., 1992) using the scikit-allel v1.3.1 library (Miles et al., 2020) in Python. As discussed by Bhatia et al. (Bhatia et al., 2013), Hudson's method performs better when the sample size per breed varies largely and demands less computational power for big datasets.

2.2.3 Admixture

The population structure analysis was done with the Admixture v1.3 software (Alexander et al., 2009). For this analysis, we used the segregating sequence variants of the pruned dataset (the same used for the phylogenetic analysis) of the 20 breeds most closely related to DSN according to their F_{ST} values. The number of animals for DSN and Holstein was reduced to 50, which were selected based on kinship. The 50 least related animals were selected based on a genomic relationship matrix (Yang et al., 2011) calculated with PLINK v2 (--make-rel) following a greedy approach starting with a randomly selected animal. Unsupervised analyses were performed for K (number of ancestral populations) ranging from 2 to 20 with 5-fold cross-validation (CV), whereof K = 4 was considered for interpretation, with the lowest CV error (CV error = 0.1448). K = 5, 6, and 7, also showed very low CV errors (Supplementary Figure S2). Those results were visualized with the CLUMPAK v1.1 software (Kopelman et al., 2015). Admixture between breeds was confirmed based with f3 statistics calculated on TreeMix v1.13 software using the *threepop* function over blocks of 10,000 variants.

2.3 Diversity analyses

2.3.1 Genomic diversity

The genomic variation within breeds was assessed as nucleotide diversity (π) (Nei and Li, 1979), and as observed (H_o) and expected (H_e) heterozygosity. H_o and H_e were estimated using vcftools v0.1.15 (Danecek et al., 2011), calculated relative to the total number of variants among all 24 breeds (34,856,428). This was done in order to remove bias on the number of segregating variants per breed, and to allow the comparison between breeds. A chi-squared test was performed between H_o and H_e for each breed using the package statsmodels v0.13.5 (Seabold and Perktold, 2010) in Python. Nucleotide diversity was calculated per window of 10 kb (π_{window}) using the library scikit-allel v1.3.1 (Miles et al., 2020) in Python. The nucleotide diversity per chromosome was calculated as the mean ($\pi_{ChrMean}$) and median ($\pi_{ChrMedian}$) of all π_{window} values of the respective chromosome, and total nucleotide diversity as the mean ($\pi_{TotMean}$) and median ($\pi_{TotMedian}$) of all π_{window} values across the whole genome. Due to the non-parametric nature of the π_{window} distributions, we also report median values in the Supplementary Materials (Supplementary Table S4). Since $\pi_{TotMean}$ and $\pi_{TotMedian}$ are highly correlated (Pearson correlation coefficient of 0.99), both values allow the same interpretation of the results. We preferred means over medians since means are widely used for nucleotide diversity in the literature, providing opportunities for comparisons. All π measures are shown as percentages, which means that final results were multiplied by 100.

2.3.2 Genomic inbreeding

Genomic inbreeding was assessed using two estimators:

(1) Excess of homozygosity (F_{Hom}) which is based on the method of moments (Li and Horvitz, 1953) was calculated as

$$F_{Hom} = \frac{\text{number of observed homozygous} - \text{number of expected homozygous}}{\text{total number of variants} - \text{number of expected homozygous}},$$

and

(2) Inbreeding coefficient F_{RoH} which is based on RoHs was estimated using BCFtools v1.9 with an assumed recombination rate of 10^{-8} per base pair (1 cM/Mb). Inbreeding was calculated for the 24 breeds that were used for diversity analysis. Allele frequency was estimated using vcftools v0.1.15 and used as an input parameter for BCFtools v1.9. RoHs were separated into five groups with the minimal lengths of 50 kb, 100 kb, 1 Mb, 2 Mb, and 4 Mb (McQuillan et al., 2008; Zhang et al., 2015; Forutan et al., 2018; Bhati et al., 2020; Dixit et al., 2020). F_{RoH} was calculated for each group as

$$F_{RoH} = \sum L_{RoH} \Big/ L_{genome},$$

where L_{RoH} is the length of a homozygous region and L_{genome} the length of the genome covered by SNPs (2,487,849,970 bp for our dataset).

2.4 Signatures of selection

2.4.1 Region-wide F_{ST} between DSN and Holstein

F_{ST} values between DSN and Holstein were compared across the whole genome in windows of 10 kb. Only windows containing at least five sequence variants were considered. The top 0.01 percentile of the F_{ST} values were selected to point to potential differences in selection signatures of those breeds.

2.4.2 Cross-population-extended haplotype homozygosity

The cross-population-extended haplotype homozygosity (XP-EHH) (Sabeti et al., 2007) was calculated between DSN and Holstein using the R package rehh v3.2.2 (Gautier et al., 2017). All segregating variants occurring in DSN or Holstein were considered. Variants were initially phased using Beagle v5.1 (Browning et al., 2018). Positive XP-EHH scores represent variants positively selected in DSN compared to Holstein, and negative scores correspond to variants positively selected in Holstein compared to DSN. *p*-values were corrected for multiple testing using Bonferroni and variants with a *p*-value<0.05 were considered as significant. The positions of neighboring significant variants were considered together as significant XP-EHH region for gene annotation.

2.4.3 RoH islands

RoH islands were defined as regions with the highest frequency of SNPs inside RoHs among DSN or Holstein.

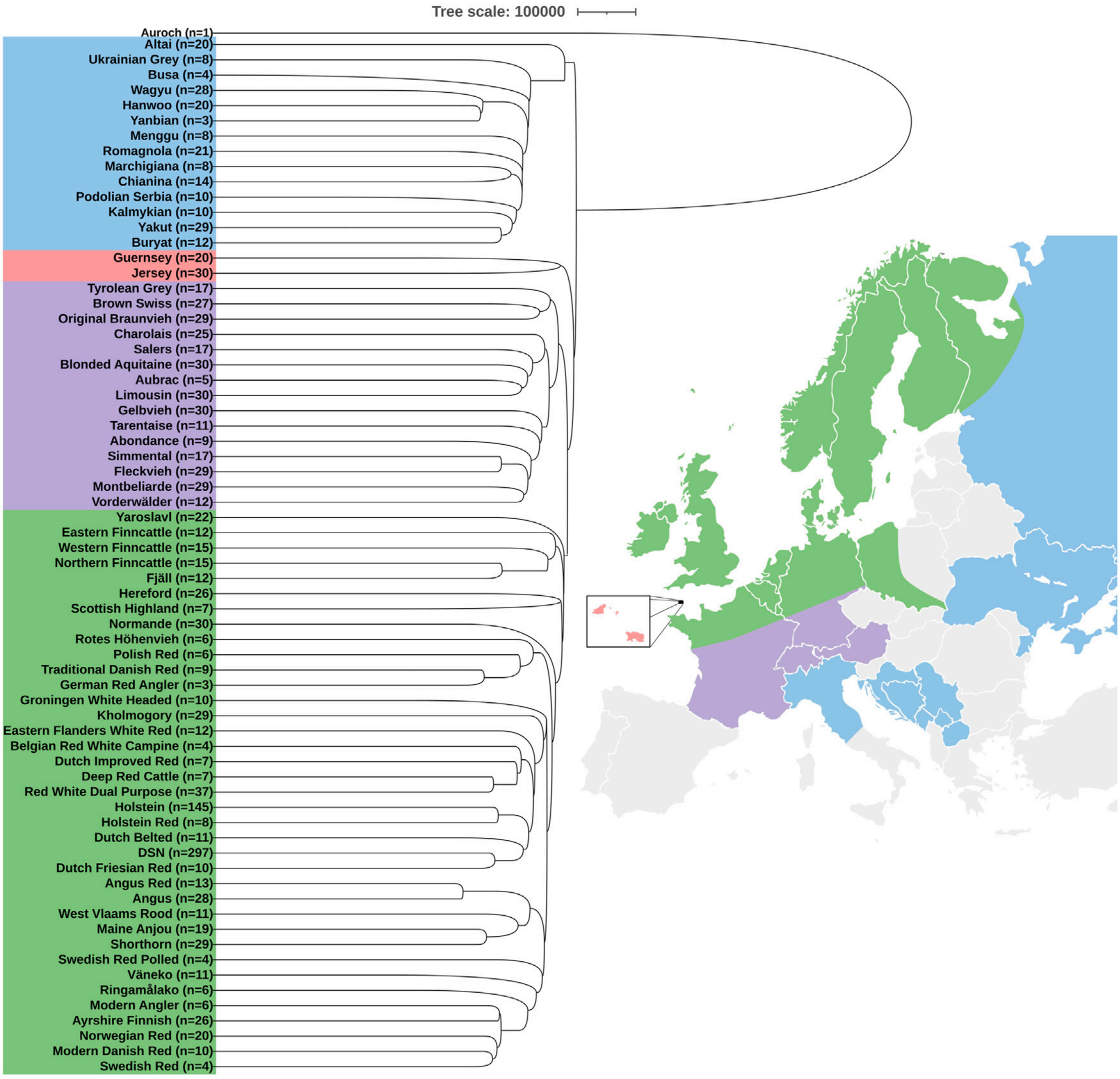


FIGURE 1
Phylogenetic tree and diversity analysis of 69 cattle breeds and Auroch as an outgroup. Colors in the phylogenetic tree represent geographical location of cattle origins: Northern Europe (green), Central Europe (violet), Jersey and Guernsey islands (red), and Eastern Europe together with Central Italy and Asia (blue). In parentheses, n is the number of animals representing the breed. Countries without any genomic breed information are highlighted in gray.

Frequency of SNP inside RoHs were calculated as the number of animals in which a SNP was reported inside a RoH, divided by the total number of animals in the breed. The threshold to define islands was taken as the top 0.05 percentile of frequencies. Afterwards, the positions of neighboring SNPs satisfying the defined percentile threshold were used to form islands.

2.4.4 Gene annotation

Regions with high difference in F_{ST} values between DSN and Holstein, significant XP-EHH regions, and RoH islands ± 250 kb flanking the start and end positions were scanned for protein-coding genes using the Ensembl database release 106 and for QTLs stored in the CattleQTLdb release 47 (Hu et al., 2022). The flanking 250 kb regions were included in our search since linkage groups and haplotype blocks can be quite large in DSN and reach, e.g., in the casein region on BTA 6, even 1 Mb (Korkuć et al., 2021). From the CattleQTLdb, QTLs and associations for “production”, “exterior”, “meat and carcass” (meat), “milk”, “health”, and “reproduction” (fertility) traits of a length <10 kb were considered. The references for the QTLs and associations were retrieved using PubMed IDs through the metapub v0.5.5 Python package. The complete list of publications is shown in Supplementary Tables S7–S9.

3 Results

3.1 Phylogenetic analysis

The phylogenetic analysis of DSN and 68 other cattle breeds including Auroch as an outgroup showed four clear clusters based on geographical origins (Figure 1). The majority of the breeds including DSN and Holstein formed a cluster of Northern European countries. Within this cluster, many Holstein Red cattle were found within the sub-cluster of Holstein cattle and *vice versa* (Supplementary Table S1), which was expected, since the coat color is the main difference. Also, animals of the breeds Modern Danish Red, Swedish Red, Norwegian Red, Ayrshire Finish, and Modern Angler were mixed with each other, showing wrong assignments or high relationship between those breeds exists (Supplementary Table S1). The cluster of Central Europe comprised all breeds from Austria, Switzerland, Southern Germany and France. Those are mainly dual-purpose breeds kept in mountainous areas. Jersey and Guernsey formed a separate cluster. The remaining cluster was formed by breeds from Eastern European countries, Central Italy, and Asian countries. This last cluster was the closest to Auroch.

We observed that DSN clustered closely to Dutch Friesian Red, Dutch Belted, Holstein, Holstein Red, Red White Dual Purpose, Deep Red Cattle, Dutch Improved Red, Belgian Red White Campine, Eastern Flanders White Red, Kholmogory, and Groningen White Headed. Those are all breeds originating from Germany, the Netherlands, and Belgium, except for Kholmogory which is from Russia, but crossbreeding with Friesian cattle was reported (Dmitriev and Ernst, 1987).

It is important to mention that the Finncattle breeds, Fjäll, Scottish Highland, Yaroslavl, and Altai clustered outside our reported clusters depending on the pruning parameters (Supplementary Figure S3). In addition, migration events were observed between an ancestor of Shorthorn and Maine Anjou to Charolais and from Traditional Danish Red to an ancestor of Polish Red and Rotes Höhenvieh (Supplementary Figure S4).

3.2 Relatedness to DSN using F_{ST}

The same breeds as in the phylogenetic tree analysis showed the closest relationship to DSN in terms of lowest F_{ST} values. Those breeds with increasing F_{ST} values from 0.032 to 0.075 are Dutch Friesian Red, Dutch Improved Red, Eastern Belgian Red White, Belgian Red White Campine, Deep Red Cattle, Eastern Flanders White Red, Holstein Red, Kholmogory, Red and White Dual Purpose, Holstein, and Dutch Belted. Breeds which have also low F_{ST} values ranging between 0.053 and 0.078, but a little bit more distant in the phylogenetic analysis, are Modern Angler, Modern Danish Red, German Red Angler, Polish Red, Norwegian Red, Eastern Finncattle, West Vlaams Rood, and Swedish Red Polled (listed in increasing order). Charolais is the only breed that clustered outside the North Europe cluster in the phylogenetic tree, but had a low F_{ST} value of 0.070 to DSN. All F_{ST} values are listed in Supplementary Table S2.

3.3 Admixture

The admixture analysis (Figure 2) at $K = 4$ corroborated the results from the phylogenetic analysis by showing a common ancestral population between DSN, Dutch Friesian Red, Dutch Belted, Kholmogory, Dutch Improved Red, Eastern Flanders White Red, and Eastern Belgian Red White (in blue)—all breeds from the Netherlands and Belgium, except for Kholmogory. Polish Red, Swedish Red Polled, Modern Angler, Norwegian Red, German Red Angler, Modern Danish Red, Charolais, Eastern Finncattle, West Vlaams Rood, and Eastern Flanders White Red appear with a common ancestry (in orange). Holstein and Holstein Red share the same ancestry (in dark blue), with a clear level of introgression in most of the breeds, including DSN. Even though Charolais is the only breed located in a different cluster in the phylogenetic tree, admixture between Charolais, DSN, and breeds closely related to DSN exists. Admixture was detected by f3 statistics, whereof Charolais is significantly admixed from DSN and all the other 19 tested populations (Supplementary Table S3).

3.4 Genomic diversity

The average value of the expected heterozygosity per individual (H_e) ranged between 9.4% in DSN and 11.9% in Yakut and the average observed heterozygosity (H_o) ranged between 9.2% in Jersey and 11.3% in Yakut (Figure 3A). Although the expected and observed heterozygosity did not significantly differ in DSN, absolutely, DSN did not have the lowest observed heterozygosity. With a few exceptions (Holstein, Norwegian Red, Blonded Aquitaine, Normande, Montbeliarde, Shorthorn, and Brown Swiss), most breeds did not show a H_o significantly lower than H_e (Supplementary Table S4).

The total average genomic diversity per individual $\pi_{TotMean}$ ranged between 0.136% in Jersey and 0.169% in Yakut (Figure 3A). In DSN, $\pi_{TotMean}$ was 0.151%. The DSN value was higher than in Holstein (0.147%) and in other breeds such as Red White Dual Purpose, Norwegian Red, Swedish Red, Ayrshire Finnish, Hereford, Angus, Normande, Montbeliarde, Shorthorn, Brown Swiss, Jersey, and Wagyu, but lower than in Modern Danish Red, Kholmogory, Charolais, Gelbvieh, Limousin, Simmental, Original Braunvieh, Fleckvieh, and Yakut (Supplementary Table S3). The values across all breeds for $\pi_{TotMean}$ and H_o ($r = 0.87$, $p = 4.5 \times 10^{-8}$) as well as between $\pi_{TotMean}$ and H_e ($r = 0.88$, $p = 1.7 \times 10^{-8}$) correlated highly significantly (Supplementary Table S5).

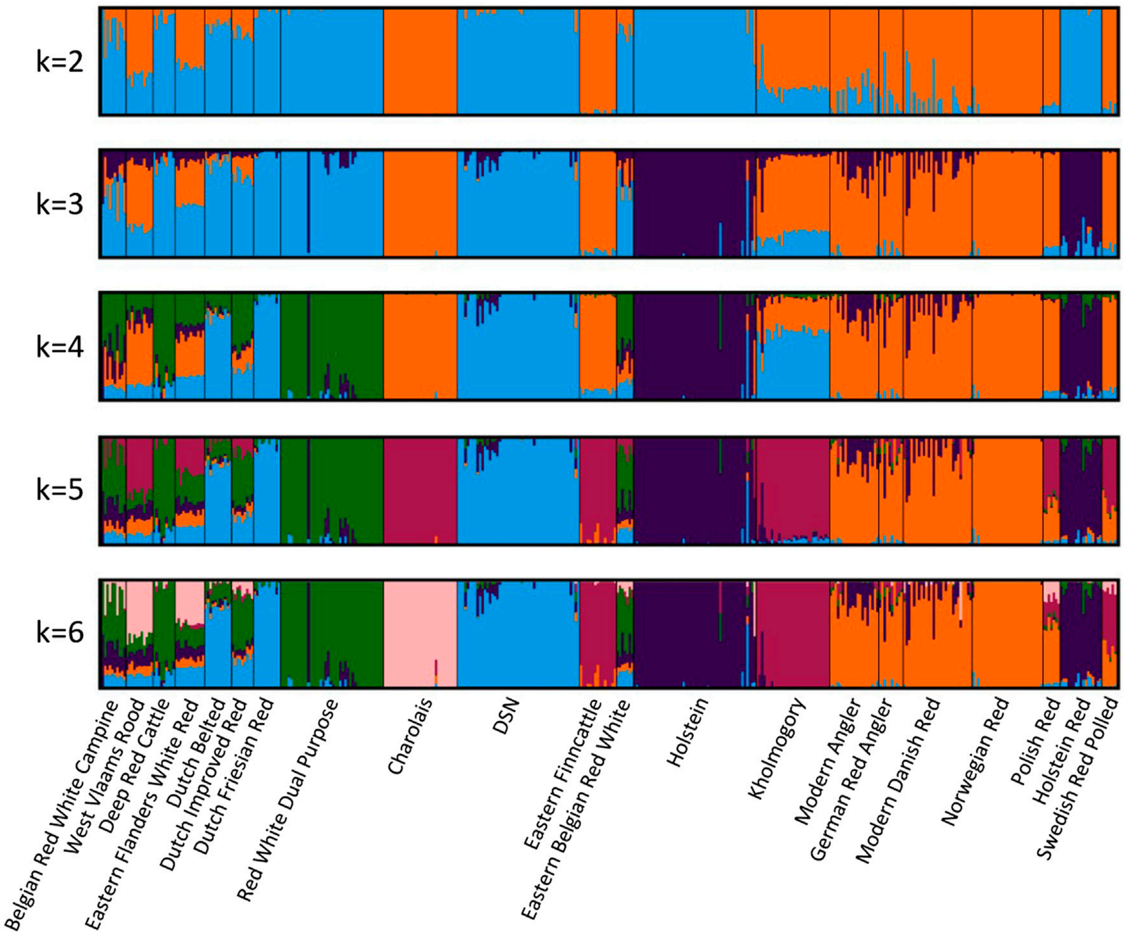


FIGURE 2
Admixture analysis between DSN and the 20 breeds closest to DSN according to the F_{ST} values. K = 4 was optimal for the 5-fold cross-validation. Colors indicate the ancestral relationship between breeds.

When looking at the nucleotide diversity per chromosome, $\pi_{ChrMean}$ values in DSN ranged between 0.134% on BTA 22 and 0.192% on BTA 23. The highest nucleotide diversity was evident in all breeds on BTA 23 in the window between 19 and 37 Mb (Figure 3B; Supplementary Table S6). This highly polymorphic region contains many protein-coding genes including the major histocompatibility complex (MHC) and the bovine leukocyte antigen-BoLA. Consistent with the total genomic nucleotide diversity, Jersey and Yakut showed the lowest and highest levels of diversity for most of the chromosomes, respectively (Figure 3B).

3.5 Genomic inbreeding

The inbreeding rate estimated as excess of homozygosity (F_{Hom}) ranged from 2.7% in DSN to 9.8% in Hereford (Figure 3C). Hence, F_{Hom} in DSN was lower than in Holstein (6.4%) and all the other breeds. As expected, commercial breeds generally showed higher inbreeding rates, as seen for example in Charolais (7.0%), Shorthorn (8.5%), Brown Swiss (9.4%), and Hereford (9.8%).

In total, 1,337,471 RoHs were detected in the 302 sequenced DSN animals with an average length of 102 kb ranging from 526 bp to 14 Mb. On average, 4,428 RoHs were detected per DSN animal. In 150 Holstein animals, in total 507,198 RoHs were found with an average length of 128 kb ranging from 851 bp to 21 Mb. The average number of RoHs per Holstein animal was 4,226. This shows the presence of longer RoHs in Holstein in comparison to DSN.

The inbreeding rate as estimated by F_{RoH} and considering all RoHs longer than 50 kb was on average 16.0% in DSN and 19.7% in Holstein (Figure 3C). From $F_{RoH>50kb}$ to $F_{RoH>2Mb}$, lowest and highest inbreeding was

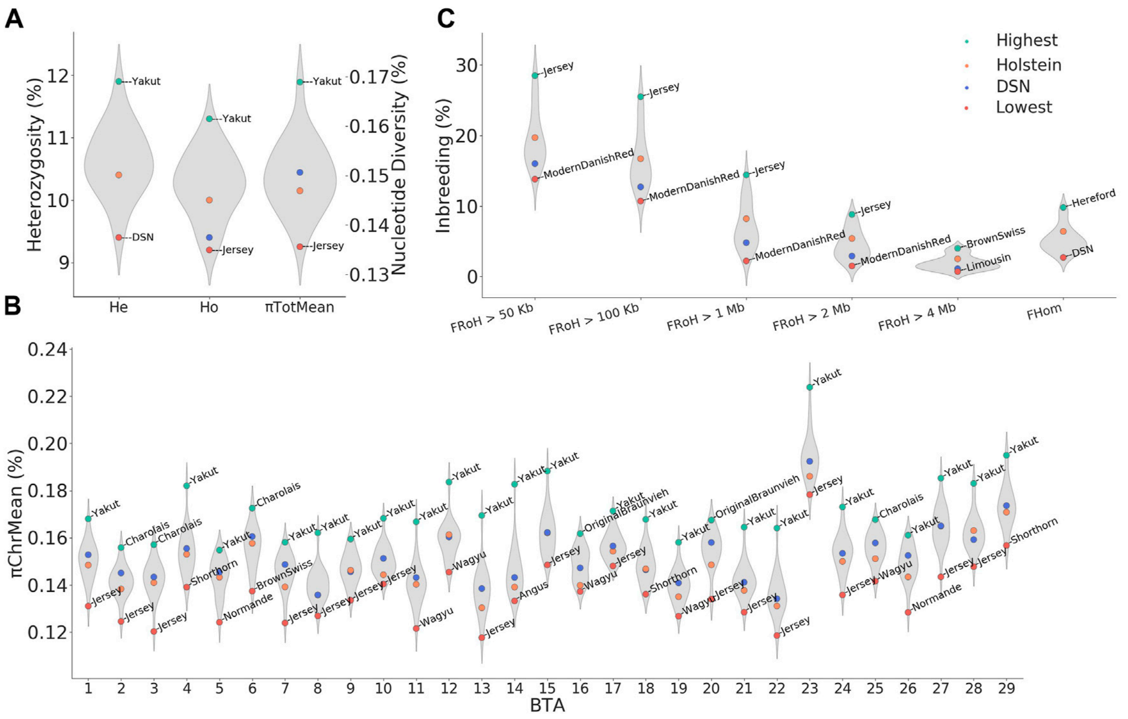


FIGURE 3
Overview of diversity measurements and inbreeding across breeds. **(A)** Average nucleotide diversity ($\pi_{TotMean}$) observed heterozygosity (H_o), andexpected heterozygosity (He) distributions, **(B)** average nucleotide diversity per chromosome ($\pi_{TotMean}$), **(C)** inbreeding defined based on runs ofhomozygosity (F_{RoH}) and excess of homozygosity (F_{Hom}). Values for DSN and Holstein are highlighted in blue and orange, respectively. For each distribution, the highest and lowest values are highlighted in green and red, respectively, including the respective breed names.

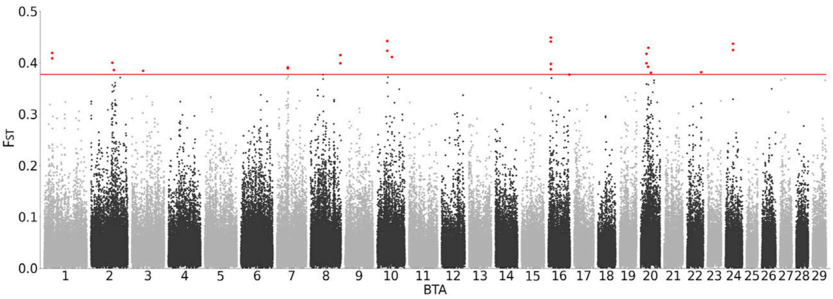


FIGURE 4
Region-wide F_{ST} between DSN and Holstein. All values were plotted in windows of 10 kb.

always observed in Modern Danish Red and Jersey, respectively, while $F_{RoH>4Mb}$ ranged from 0.7% in Limousin to 4.0% in Brown Swiss (Figure 3C). $F_{RoH>1Mb}$ and $F_{RoH>2Mb}$ showed similar results to those detected in F_{Hom} (Figure 3C). Significant Pearson's correlations of 0.60 ($p = 1.9 \times 10^{-3}$) and 0.62 ($p = 1.3 \times 10^{-3}$) were found between F_{Hom} and $F_{RoH>1Mb}$, and F_{Hom} and $F_{RoH>2Mb}$, respectively (Supplementary Table S5).

3.6 Region-wide F_{ST} between DSN and Holstein

The analysis of F_{ST} values as a measure for the genomic diversity between DSN and Holstein revealed regions of high differentiation between those populations (Figure 4). Peaks of high F_{ST} values for windows of 10 kb length were identified on BTA 1, 2, 3, 7, 8, 10, 16, 20, 22 and 24. F_{ST} values of 25 10 kb-

TABLE 2 Positional candidate genes in regions with the highest F_{ST} values between DSN and Holstein. Genes in bold are located directly inside the windows with the highest F_{ST} values. All other genes are located ± 250 kb from the start and end positions of the windows with the highest F_{ST} values. Consecutive windows displaying the same genes are shown together, this is the case for windows in BTA 1, 7, 8, and 10, whereof length is 20 kb.

Location	SNP count	Length (kb)	Genes ± 250 kb
1 : 29,280,001–29,300,000	279	20	1,4-alpha-glucan branching enzyme 1 (***GBE1***), *ENSBTAG00000008359*
2 : 78,690,001–78,700,000	91	10	Glycophorin C (*GYPC*)
2 : 84,980,001–84,990,000	88	10	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2 (***HECW2***), coiled-coil domain containing 150 (*CCDC150*)
3 : 41,600,001–41,610,000	97	10	Olfactomedin 3 (*OLFM3*), *ENSBTAG00000051863*
7 : 41,060,001–41,080,000	346	20	Germinal center associated signaling and motility like (*GCSAML*), olfactory receptor family 2 subfamily C member 3 (*OR2C3*), and 3B (*OR2C3B*), subfamily G member 2 (*OR2G2*), 3 (*OR2G3*), and 27 *(OR2G27*), subfamily W member 3 (*OR2W3*), and 3D (*OR2W3D*), subfamily AO member 1 (*OR2AO1*), subfamily T member 54 (*OR2T54*), family 5 subfamily AE member 3 (*OR5AE3*), and 4 (*OR5AE4*), family 6 subfamily F member 1 (*OR6F1*), subfamily AA member 1 (*OR6AA1*), subfamily AN member 1 (*OR6AN1*), family 9 subfamily E member 2 (*OR9E2),* family 11 subfamily L member 1 (*OR11L1*), family 14 subfamily P member 2 (*OR14P2*), tripartite motif containing 58 (*TRIM58*), *ENSBTAG00000030735*
8 : 110,900,001–110,920,000	265	20	Doublecortin domain containing 2C (***DCDC2C***), gelsolin (*GSN*), stomatin (*STOM*), allantoicase (*ALLC*), collectin subfamily member 11 (*COLEC11*), ribosomal protein S7 (*RPS7*), ribonuclease H1 (*RNASEH1*), acireductone dioxygenase 1 (*ADI1*), trafficking protein particle complex subunit 12 (*TRAPPC12*), EARP complex and GARP complex interacting protein 1 (*EIPR1*), *ENSBTAG00000049154*
10 : 38,480,001–38,500,000	206	20	Ubiquitin protein ligase E3 component n-recognin 1 (*UBR1*), transmembrane protein 62 (*TMEM62*), cyclin D1 binding protein 1 (*CCNDBP1*), erythrocyte membrane protein band 4.2 (*EPB42*), *ENSBTAG00000046363*
10 : 55,600,001–55,610,000	177	10	Unc-13 homolog C (*UNC13C*)
20 : 27,500,001–27,510,000	77	10	ISL LIM homeobox 1 (*ISL1*)
20 : 28,610,001–28,620,000	119	10	Poly (ADP-ribose) polymerase family member 8 (*PARP8*), embigin (*EMB*)
20 : 37,390,001–37,400,000	101	10	Ciliosis and planar polarity effector 1 (*CPLANE1*), NIPBL cohesin loading factor (*NIPBL*), *ENSBTAG00000050782,* solute carrier family 1 member 3 (*SLC1A3*)
22 : 52,820,001–52,830,000	105	10	Coiled-coil domain containing 12 (*CCDC12),* parathyroid hormone 1 receptor (*PTH1R*), myosin light chain 3 (*MYL3*), serine protease 42 (*PRSS42P*), *ENSBTAG00000037821, ENSBTAG00000052304, ENSBTAG00000049544, ENSBTAG00000050911,* serine protease 45 (*PRSS45*), *ENSBTAG00000038616, ENSBTAG00000005019*, transmembrane inner ear (*TMIE*), ALS2 C-terminal like (*ALS2CL),* leucine rich repeat containing 2 (*LRRC2*), teratocarcinoma-derived growth factor 1 (*TDGF1*), receptor transporter protein 3 (*RTP3*), lactotransferrin (*LTF*), C-C motif chemokine receptor like 2 (*CCRL2*), receptor 5 (*CCR5*), and 2 (*CCR2*)

regions were above the threshold of 0.38 which represents the 0.01 percentile of the F_{ST} distribution. Since the average F_{ST} between DSN and Holstein was 0.069, most variants showed only low F_{ST} values.

The genes encoding 1,4-alpha-glucan branching enzyme 1 (*GBE1*), HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2 (*HECW2*), and doublecortin domain containing 2C (*DCDC2C*) are located inside the 10 kb-regions on BTA 1, 2, and 8, respectively, with F_{ST} values between DSN and Holstein above the threshold. Additional 71 genes are located 250 kb up- or downstream of the start and end positions of the 25 detected 10 kb-regions above the threshold (Table 2). Regions on BTA 16 and 24 do not contain protein-coding genes. Annotation on QTLs and associations of all regions from CattleQTLdb are shown in Supplementary Table S7.

3.7 Cross-population-extended haplotype homozygosity

The analysis of XP-EHH identified 140 variants positively selected in DSN in comparison to Holstein and 80 variants positively selected in Holstein in comparison to DSN (Figure 5). Those variants are located in DSN in four regions on BTA 5, 12, 18, and 29, and in Holstein in five regions on BTA 2, 8, 10, 18, and 23 (Table 3). The highest number of significant variants was detected on BTA 12 in DSN (124 variants) and on BTA 18 in Holstein (57 variants). 73% of the significant variants were intronic, while 26% were intergenic and 1% downstream of genes.

Seven genes are located directly within the identified XP-EHH regions, among them four in DSN and three in Holstein. When considering genes located 250 kb up- or downstream of

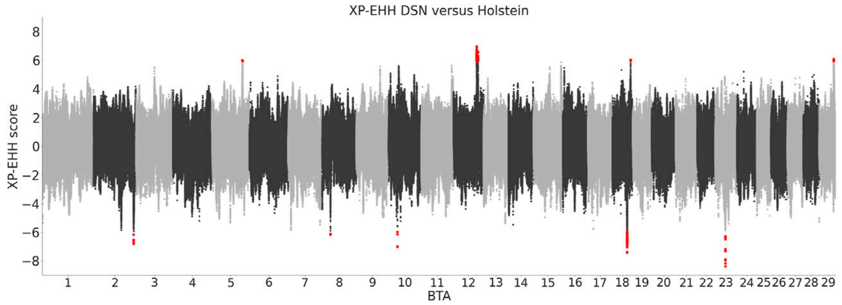


FIGURE 5
XP-EHH scores between DSN and Holstein. Positive scores represent variants positively selected in DSN in comparison to Holstein, and negative scores represent variants positively selected in Holstein in comparison to DSN. Significant scores are shown in red.

the start and end positions of the regions, 110 genes are detected, 49 in DSN, and 61 in Holstein. Furthermore, no overlaps were detected between regions with high F_{ST} and XP-EHH regions. Annotation of genes in QTLs and associations of all regions from CattleQTLdb are shown in Supplementary Table S8.

3.8 RoH islands

Based on the frequency of SNPs inside RoHs, 21 RoH islands were detected in DSN and 19 in Holstein. Supplementary Figure S5 shows the frequency of SNPs inside RoHs for DSN and Holstein, and the threshold used in each case to define islands (0.66 in DSN and 0.61 in Holstein). Regions on BTA 1, 4, 14, and 18, where RoH islands were detected in both breeds in the same location, likely contributed essentially to selection (Table 4).

In total, 26 genes were found inside all RoH islands in DSN. For Holstein, 17 genes were located inside all RoH islands. Between DSN and Holstein, nine genes overlapped with the RoH islands on BTA 4, 14, and 18. Genes occurring 250 kb up- or downstream from start and end positions of RoH islands are shown in Supplementary Table S9.

No exact overlap was detected between RoH islands and regions with high F_{ST} or XP-EHH regions, neither in DSN nor in Holstein. In DSN, the smallest distance of a RoH island to an XP-EHH region was 3.8 Mb on BTA 18 ($R^2 = 0.001$, D' = 0.12), and to an F_{ST} window 7.5 Mb on BTA 10 ($R^2 = 0.001$, D' = 1). In Holstein, the smallest distance of a RoH island to an XP-EHH region was 3.0 Mb on BTA 20 ($R^2 = 0.03$, D' = 0.22) and to an F_{ST} window 4.6 Mb on BTA 8 ($R^2 = 0.03$, D' = 0.23).

4 Discussion

4.1 Relationship analyses

Considering the origin of DSN in the North Sea region, our expectation of cattle breeds from the same region such as Dutch Friesian Red, Holstein, and other breeds share a common ancestry with DSN was confirmed. But interestingly, Holstein was not the closest related breed to DSN, neither in F_{ST} nor phylogenetic analyses, despite their shared history. Asian and Eastern European countries, and Central Italy showed the closest cluster to Auroch, indicating the most ancient origin. Cattle domestication in Europe is believed to have started in Italy, with further migration of those cattle to Central and Northern Europe (Felius et al., 2014), which is consistent with the clusters of breeds in our phylogenetic tree. In addition, our clusters are also consistent with clusters previously reported by other genetic studies (Felius et al., 2011) that are also based on WGS data of cattle (Dutta et al., 2020). Since pruning parameters slightly affect the tree construction, and considering the complexity of the development of the cattle breeds in Europe (Felius et al., 2014), we cannot entirely explain how much the true relationship between breeds differs from our or reported findings.

We identified an F_{ST} value of 0.069 between DSN and Holstein. This value is consistent with the estimate of 0.068 which has been reported using 261 and 4,654 animals from Illumina BovineSNP50 Beadchip data, respectively (Naderi et al., 2020). Generally, F_{ST} values support our phylogenetic findings. Nevertheless, small differences were seen in the order of closest related breeds to DSN based on F_{ST} values, and breeds inside clusters closely related to DSN in the phylogenetic tree. This may be caused by the fact that the UPGMA algorithm used the average distance between all investigated animals at once to build the phylogenetic tree, while F_{ST} values were calculated pairwise between DSN and each breed separately. The latter might slightly bias the analysis towards those two breeds. In addition, the data used for the phylogenetic tree was pruned, while F_{ST} values were calculated using only unpruned variants segregating between two investigated breeds.

Although Charolais was not located inside the Northern European cluster in the phylogenetic tree, it showed a low F_{ST} value of 0.070 with DSN. Such a low F_{ST} value of 0.074 was also reported between Charolais and Holstein (Kelleher et al., 2017). The

TABLE 3 Positional candidate genes in regions with significant XP-EHH regions between DSN and Holstein. Genes in bold are located directly within the regions. All other genes are located ± 250 kb from the start and end positions of the regions.

Location	Length (kb)	SNP count	Breed whereof positively selected	Genes ± 250 kb
2 : 130,479,391–130,479,547	0.156	6	Holstein	EPH receptor A8 (*EPHA8*), Zinc finger and BTB domain containing 40 (*ZBTB40*), Wnt family member 4 (*WNT4*)
8 : 23,020,644–23,020,649	0.005	2	Holstein	*ENSBTAG00000048891*, Interferon-tau-like (*IFN-TAU*), *ENSBTAG00000054099, ENSBTAG00000055152*, Kelch like family member 9 (*KLHL9*), Interferon alpha G (*IFNAG*), *ENSBTAG00000050194, ENSBTAG00000052859, ENSBTAG00000048428, ENSBTAG00000051881, ENSBTAG00000053037, ENSBTAG00000053413*, Interferon beta 3 (*IFNB3*), *ENSBTAG00000046967*
10 : 23,771,505–23,771,788	0.283	4	Holstein	*ENSBTAG00000051554, ENSBTAG00000048374, ENSBTAG00000052580,* ENSBTAG00000048874, T cell receptor alpha variable 24 (*TRAV24*), *ENSBTAG00000052314*
18 : 51,426,960–51,445,790	18.830	57	Holstein	*ENSBTAG00000054584*, Glutamate ionotropic receptor kainate type subunit 5 (*GRIK5*), ATPase Na+/K+ transporting subunit alpha 3 (*ATP1A3*), Rab acceptor 1 (*RABAC1*), *ENSBTAG00000053222*, Rho guanine nucleotide exchange factor 1 (*ARHGEF1*), CD79a molecule (*CD79A*), Ribosomal protein S19 *(RPS19*), DMRT like family C2 (*DMRTC2*), LY6/PLAUR domain containing 4 (*LYPD4*), *ENSBTAG00000006859*, ***ENSBTAG00000049346***, *ENSBTAG00000052343, ENSBTAG00000054156*, C-X-C motif chemokine ligand 17 (*CXCL17*), CD177 molecule (*CD177*), Testis expressed 101 (*TEX101*), Binder of sperm 1 (*BSP1*), 3 (*BSP3*), and 5 (*BSP5*), *ENSBTAG00000049614*, LY6/PLAUR domain containing 3 (*LYPD3*), Pleckstrin homology like domain family B member 3 (*PHLDB3*)
23 : 26,112,509–26,271,305	158.796	11	Holstein	Butyrophilin like 2 (*BTNL2*), *ENSBTAG00000034945, ENSBTAG00000007618,* ***ENSBTAG00000026163, ENSBTAG00000050817***
5 : 99,140,148–99,322,102	181.954	4	DSN	Serine/threonine/tyrosine kinase 1 (*STYK1*), Mago homolog B exon junction complex subunit (*MAGOHB*), *ENSBTAG00000009252, ENSBTAG00000052865, ENSBTAG00000046268,* ***ENSBTAG00000049367***, ***ENSBTAG00000054018, ENSBTAG00000054633***, *ENSBTAG00000052486, ENSBTAG00000052514, ENSBTAG00000050324, ENSBTAG00000022861, ENSBTAG00000051183, ENSBTAG00000049823, ENSBTAG00000052658, ENSBTAG00000038843*
12 : 69,770,246–71,722,914	1,952.668	124	DSN	Multidrug resistance-associated protein 4 (*LOC515333*), ***ENSBTAG00000047383, ENSBTAG00000046041, ENSBTAG00000049836***
18 : 61,391,310–61,398,262	6.952	3	DSN	*ENSBTAG00000000336, ENSBTAG00000009171, ENSBTAG00000015061, ENSBTAG00000014328, ENSBTAG00000054918, ENSBTAG00000013345, ENSBTAG00000009364, ENSBTAG00000015987, ENSBTAG00000051856, ENSBTAG00000046961, ENSBTAG00000051149, ENSBTAG00000030416, ENSBTAG00000015139, ENSBTAG00000018152*, Protein kinase C gamma (*PRKCG*), Calcium voltage-gated channel auxiliary subunit gamma 7 (*CACNG7*), and 8 (*CACNG8*)
29 : 38,542,360–38,548,547	6.187	9	DSN	*ENSBTAG00000051614, ENSBTAG00000039970, ENSBTAG00000050440, ENSBTAG00000040340*, Pregnancy-associated glycoprotein 1 (*PAG1*), *ENSBTAG00000048202, ENSBTAG00000054803, ENSBTAG00000051196*

admixture and f3 analyses provided evidence for some admixture between Charolais and some other breeds from Northern Europe. Additionally, a migration event was observed between an ancestor of two breeds from the Northern European cluster, Shorthorn and Maine Anjou, to Charolais, which might explain the admixture results. An influence of shorthorn on Charolais –or of Durham cattle, ancestor of Shorthorn and Maine Anjou – has been reported before from the breeding history. The two breeds were separated by establishing independent herd books only in 1890 (Felius et al., 2014).

High admixture was observed between Modern Angler, Modern Danish Red, German Red Angler, and Norwegian

TABLE 4 Location of RoH islands and genes inside RoH islands in DSN and in Holstein. Genes in bold are common between breeds.

BTA	RoH island in DSN		RoH island in Holstein	
	Location	Genes	Location	Genes
4	76,848,259–76,993,327	Transmembrane p24 trafficking protein 4 (***TMED4***), DEAD-box helicase 56 (***DDX56***), NPC1 like intracellular cholesterol transporter 1 (***NPC1L1***), NudC domain containing 3 (***NUDCD3***)	76,847,893–76,984,479	Transmembrane p24 trafficking protein 4 (***TMED4***), DEAD-box helicase 56 (***DDX56***), NPC1 like intracellular cholesterol transporter 1 (***NPC1L1***), NudC domain containing 3 (***NUDCD3***)
6	35,779,310–35,831,557	Family with sequence similarity 13 member A (*FAM13A*)		
8			106,068,712–106,290,180 (with gaps, see Supplementary Table S9)	Astrotactin 2 (*ASTN2*)
10	30,721,933–30,889,605	Diphthamine biosynthesis 6 (*DPH6*)		
13			437,055–501,916 (with gaps, see Supplementary Table S9)	*ENSBTAG00000024139, ENSBTAG00000051498,* olfactory receptor family 4 subfamily C member 27 (*OR4C27*)
14	22,768,759–23,297,958 (with gaps, see Supplementary Table S9)	XK related 4 (***XKR4***), transmembrane protein 68 (***TMEM68***), trimethylguanosine synthase 1 (***TGS1***)**,** LYN proto-onco, Src family tyrosine kinase (***LYN***)*,* ribosomal protein S20 (*RPS20*)	22,768,535–23,225,387 (with gaps, see Supplementary Table S9)	XK related 4 (***XKR4***), transmembrane protein 68 (***TMEM68***), trimethylguanosine synthase 1 (***TGS1***)**,** LYN proto-onco, Src family tyrosine kinase (***LYN***)
	31,316,339–31,538,677 (with gaps, see Supplementary Table S9)	Centrosome and spindle pole associated protein 1 (*CSPP1*)*,* ADP ribosylation factor guanine nucleotide exchange factor 1 (*ARFGEF1*)*,* Carboxypeptidase A6 (*CPA6*)		
16	40,659,742–41,948,518 (with gaps, see Supplementary Table S9)	TNF superfamily member 18 (*TNFSF18*), mitofusin 2 (*MFN2*), procollagen-lysine,2-oxoglutarate 5-dioxygenase 1 (*PLOD1*), angiotensin II receptor associated protein (*AGTRAP*)		
17	60,987,746–61,013,323	LIM homeobox 5 (*LHX5*), serine dehydratase like (*SDSL*)		
18	14,422,668–14,490,133	Ankyrin repeat domain 11 (*ANKRD11*), SPG7 matrix AAA peptidase subunit, paraplegin (***SPG7***)	14,482,261–14,525,853	SPG7 matrix AAA peptidase subunit, paraplegin (***SPG7***), ribosomal protein L13 (*RPL13*), copine 7 (*CPNE7*), dipeptidase 1 (*DPEP1*)
	57,506,154–57,570,852 (with gaps, see Supplementary Table S9)	Zinc finger protein 175 (*ZNF175*), *ENSBTAG00000023365, ENSBTAG00000045880*		
26			9,558,523–9,638,107	Phosphatase and tensin homolog (*PTEN*)

Red, causing wrong assignments between those breeds in our phylogenetic analysis. This admixture, however, is consistent with recent findings (Schmidtmann et al., 2021).

Finally, we have to mention that every relationship study is dependent on the number of animals per breed, their kinship, and how good the given animals represent the breed. In our study, the number of animals per breed varied from 5 to 302, for diversity analyses from 28 to 302 sticking to FAO guidelines (FAO, 2011). We had no information on the relatedness of individuals and how good the animals represent each breed.

4.2 Genomic diversity

The average total nucleotide diversity (π_{TotMean}) of DSN was similar to other breeds from Northern Europe, but slightly above average indicating good population management. Eastern European and Asian breeds showed higher π_{TotMean} values than Northern and Central European breeds, likely due to less intensive breeding programs, genetic bottlenecks, or founder effects in comparison to the other breeds (Felius et al., 2014). The highest average chromosomal nucleotide diversity (π_{ChrMean}), which was found on BTA 23 at 19–37 Mb, can be attributed to the highly polymorphic region of the MHC. This high diversity around the MHC is observed in all mammals (Shiina et al., 2017).

Even though there was a high correlation between $\pi_{TotMean}$ and H_o and H_e, DSN showed the lowest H_e and one of the lowest H_o in comparison to other breeds, differently from its $\pi_{TotMean}$ result. One reason for this discrepancy might be the large number of 302 DSN animals in comparison to other breeds. However, this hypothesis was discarded since no correlation was detected between the number of animals per breed and $\pi_{TotMean}$. Furthermore, $\pi_{TotMean}$ calculated for either 50 (0.1513%) or 302 DSN animals (0.1506%) which was almost identical. Nevertheless, no statistical difference was detected between H_o and H_e for DSN. The high nucleotide diversity and low heterozygosity likely results from the small herd size of DSN with 2,452 cows and only 36 breeding bulls (TGRDEU, 2021).

The breeds Jersey and Yakut showed the lowest and highest levels of diversity, respectively. Jersey cattle, which were originally restricted to the Jersey island, suffer from high inbreeding and low genetic diversity rates, which is seen even among animals not kept on the islands anymore (Huson et al., 2020). Lower genetic diversity is expected for insular than for continental populations (Frankham, 1997). Yakut, in contrast, is an ancient breed from Siberia, known for its adaptation to extreme low temperatures. The reported high nucleotide diversity of 0.173% in this breed might be due to lower artificial selection and a higher effective population size of the ancestral Asian taurine in comparison to European cattle (Weldenegodguad et al., 2019).

4.3 Genomic inbreeding

Inbreeding measured by the homozygosity index F_{Hom} was generally higher for intensively selected breeds such as Hereford, Brown Swiss, Shorthorn, Charolais, and Holstein. Jersey also showed particular high inbreeding. DSN had the lowest excess of homozygosity index, which demonstrates its good management as a genetic resource. This shows that in a local breed with only about 2,500 animals, diversity can be maintained through a breeding scheme that aims at less intensive selection for production traits (Gutiérrez-Reinoso et al., 2022), which is strongly different from Holstein, for example, where millions of animals are kept. In our analysis, Holstein cattle were available from four countries, the United States, Denmark, Germany, and the Netherlands, which probably led to slightly higher diversity rates than having only Holsteins from one country.

The same pattern for inbreeding was observed when using F_{RoH} as indicated by the high correlation rates between inbreeding metrices. High correlation between F_{Hom} and F_{RoH} was also observed in the literature (Mastrangelo et al., 2016). Our F_{RoH} values were consistent with those previously reported, for example, for Original Braunvieh ($F_{RoH>50\ kb}$ = 14.58%) (Bhati et al., 2020), Modern Danish Red ($F_{RoH>10\ kb}$ = 11.84%), Holstein ($F_{RoH>10\ kb}$ = 18.67%), and Jersey ($F_{RoH>10\ kb}$ = 24.23%) (Zhang et al., 2015). Another aspect is the length of regions of homozygosity which reflects recent and ancient inbreeding. Longer RoHs indicate more recent inbreeding, while shorter RoHs indicate ancient inbreeding (Makanjuola et al., 2020). Moreover, recent inbreeding shows higher detrimental inbreeding depression effects (Makanjuola et al., 2020). Commercial breeds show more often long RoHs (Marras et al., 2015). This was the case for Hereford and Brown Swiss. Nevertheless, non-commercial breeds such as Jersey, Wagyu, and even Yakut also showed high average RoH lengths and high $F_{RoH>4\ Mb}$ (Supplementary Table S4). DSN showed fewer longer RoHs, which were generally shorter, than in other breeds.

4.4 Signatures of selection

Signatures of selection were particularly examined in DSN and Holstein cattle. Considering DSN as a dual-purpose breed for milk and meat production, and Holstein as a high-yielding dairy breed, differentiated regions are expected to contain genes associated with traits influencing meat production, carcass, body conformation, and milk production.

4.4.1 Region-wide F_{ST} between DSN and Holstein

In total, we identified 25 high differentiating F_{ST} windows likely reflecting signatures of selection in DSN when comparing to Holstein. Two regions on BTA 20 and 10 have been detected previously (Naderi et al., 2020). All other regions were novel. Inside the top high differentiating F_{ST} regions, the three genes *GBE1*, *HECW2*, *DCDC2C* were located. *GBE1* on BTA 1 was described in literature as responsible for the production of the glycogen branching enzyme, therefore, being involved in the carbohydrate and glycogen metabolisms. Glycogen is an important short-term energy storage molecule in the muscle. *GBE1* was also detected in signatures of selection using RoHs in U.S. Holstein cattle (Kim et al., 2013). Its deficiency has also been associated with glycogen storage diseases and stillbirths in humans, cattle, and equines (Ward et al., 2004; Lee et al., 2011; Lorenz et al., 2011; Almodóvar-Payá et al., 2020). Furthermore, the F_{ST} region where *GBE1* was located overlapped with two QTLs from CattleQTLdb for production traits (Crispim et al., 2015; Hamidi Hay and Roberts, 2017), growth and longevity (Supplementary Table S7). *HECW2* on BTA 2 is responsible for protein ubiquitination. The gene was described as a candidate gene for milking speed in French Holstein (Marete A. et al., 2018b), aging and angiogenesis in humans (Rotin and Kumar, 2009; Walter et al., 2011; Choi et al., 2016; Berko et al., 2017). *DCDC2C* on BTA 8 was suggested to be associated with sperm formation in cattle (Wu et al., 2020), structural defects in cilia in sperm (Jumeau et al., 2017) and in cilia length in sensorial cells in humans (Grati et al., 2015). Moreover, the same region on BTA 8 presented a QTL for meat in the CattleQTLdb (McClure et al., 2012). Those genes point for the differentiation between signatures of selection between DSN and Holstein, including candidates for meat, milk, production, and fertility traits.

In addition to genes located directly inside the topmost differentiated genomic regions between DSN and Holstein, genes within 250 kb up- or downstream were analyzed. *GYPC* on BTA 2 has been previously reported as a candidate gene for body length (Vanvanhossou et al., 2020) and subclinical ketosis in Holstein (Soares et al., 2021), while *CCDC150* was reported as candidate for milk and fat yield in Nordic Holstein cattle (Cai et al., 2020). BTA 7 showed a series of olfactory receptor genes. *GCSAML,* a germinal center associated signaling and motility like gene of mature B lymphocytes, was reported to be significantly down-regulated in Holstein cows under heat-stress conditions (Kim et al., 2021). Mammalian olfactory receptors are encoded by the largest mammalian multigene family, containing 881 genes on 26 chromosomes. Studies suggest physiological and behavior aspects of variation of olfactory receptor genes, e.g., associated to appetite regulation in livestock (Connor et al., 2018), which influences uptake of nutrients required for milk and meat production. This corroborates the idea of DSN well adapted to grazing with high roughage feed intake. In the BTA 7 region, QTLs for milk and meat production were found in the CattleQTLdb (Daetwyler et al., 2008; McClure et al., 2010; Marete A. et al., 2018b).

Other interesting genes near topmost differentiated genomic regions between DSN and Holstein are *CCNDBP1* on BTA 10, a candidate for skeletal myogenesis (Huang et al., 2016); *UNC13C* on BTA 10 a candidate for feed efficiency (Freua et al., 2016); *EMB* on BTA 20 a candidate for mammary gland tissue development (Butty et al., 2017); *NIPBL* on BTA 20 a candidate for growth (de Simoni Gouveia et al., 2017; Wang et al., 2022) and previously detected as positively selected in German Holstein in the study between DSN and German Holstein using Illumina BovineSNP50 Beadchip (Naderi et al., 2020); and *ALS2CL*, *LRRC2*, and *TDGF1* on BTA 22, which are candidates for milk production (Ibeagha-Awemu et al., 2016) and fertility (Wei et al., 2017; Tríbulo et al., 2018), and previously detected in highly differentiated regions between recent populations of Dutch Frisian and Holstein (Hulsegge et al., 2022). In the region on BTA 22, the gene *LTF* (lactotransferrin) was found which is a major iron-binding protein in milk and body secretions of bovine (Rejman et al., 1989; Pierce et al., 1991) with an antimicrobial activity (Bellamy et al., 1992). Furthermore, *LTF* was reported to influence casein yield (Cecchinato et al., 2014). Although the region on BTA 24 did not contain any gene, the same region was associated with endoparasite resistance in DSN (May et al., 2019).

4.4.2 Cross-population-extended haplotype homozygosity

Out of the four regions positively under selection in DSN, two are novel and two had been previously reported (Naderi et al., 2020). The region on BTA 12 has been reported before as positively selected in DSN in a study between DSN and German Holstein using Illumina BovineSNP50 Beadchip (Naderi et al., 2020). In this region, *LOC515333* resides, a novel gene that has been so far annotated as coding the multidrug resistance-associated protein 4 (NCBI gene ID 515333). In cattle, this protein was reported to influence fertility traits since it is involved in the transport of prostaglandins and the regulation of oxytocin (Lacroix-Pépin et al., 2011). The XP-EHH region on BTA 18 which is close to the RoH island on the same chromosome also corroborates previous findings (Naderi et al., 2020). This region contains for instance *CACNG7,* a gene that was reported as a candidate gene for feed efficiency in Nellore cattle (Olivieri et al., 2016).

The regions on BTA 5 and 29 are novel. The region on BTA 5 has been repeatedly associated with milk production (Olsen et al., 2002; Bennewitz et al., 2003; Meredith et al., 2012; Rutten et al., 2013; Jiang et al., 2019). This region contains, for example, the gene *STYK1*, which has been reported as a candidate gene for heat stress response demonstrated through milk fatty acids alterations in German Holstein (Bohlouli et al., 2022). The region on BTA 29 contains the well described gene *PGA1* (Xie et al., 1995; Klisch et al., 2005; López-Gatius et al., 2007), encoding the pregnancy associated glycoprotein-1, which is expressed in the placenta where it is crucial for a healthy gestation in cattle.

For Holstein, regions on 5 chromosomes were found. Only, the region on BTA 2 corroborates the previous study comparing DSN and German Holstein (Naderi et al., 2020). This is likely due to the different Holstein populations used in the different studies. Naderi *et al.* used Holstein cattle from Germany, while the current study used Holstein cattle from the 1000 Bull Genomes project (Hayes and Daetwyler, 2019) which originate from four different countries. The selection region on BTA 2, for instance, contains the *WNT4* gene, which affects ovulation (Tríbulo et al., 2018), neurogenesis and embryogenesis in cattle (Gao et al., 2016).

The four other selection regions detected in Holstein contain genes such as *IFNB3* on BTA 8, reported with evidence of inhibition against the bovine herspesvirus type 1 (da Silva et al., 2012), *TRAV24* on BTA 10 which is T-cell receptor, *BSP1*, *BSP2*, and *BSP3* on BTA 18 which bind the sperm proteins 1,3 and 5, respectively (D'Amours et al., 2012), and *BTNL2* on BTA 23, which encodes butyrophilin (Afrache et al., 2012), as part of the immunoglobulin superfamily of transmembrane proteins in the MHC.

The absence of an overlap between region-wide F_{ST} and XP-EHH analyses is due to differences in the two methods, which are complementary. F_{ST} values were calculated based on allele frequencies, while XP-EHH values were calculated based on the decay of haplotype homozygosity, reflecting more recent selection signatures (Sabeti et al., 2007).

4.4.3 RoH islands

When signatures of selection were analyzed within breeds, nine genes were detected by RoH islands in both DSN and Holstein. Those nine genes are located on BTA 4, 14, and 18. The region on BTA 4 has been reported for mastitis resistance (Rupp and Boichard, 2003), growth (Lu et al., 2013), fertility (Grigoletto et al., 2020b), and primarly milk (Ibeagha-Awemu et al., 2016; Sanchez et al., 2017, 2019; Van Den Berg et al., 2020), while the region on BTA 14 was primarly reported for meat (Bolormaa et al., 2011; Neto et al., 2012; Saatchi et al., 2014; Sharma et al., 2014; Ali et al., 2015; Song et al., 2016; Kim et al., 2017; Akanno et al., 2018; Grigoletto et al., 2020a; Srikanth et al., 2020; Wang et al., 2020; Rezende et al., 2021), production (Snelling et al., 2010; Lu et al., 2013; Martínez et al., 2014; Saatchi et al., 2014; Akanno et al., 2018; Zhong et al., 2019; Zhang et al., 2020), and exterior traits (Pryce et al., 2011; Pausch et al., 2016; Wu et al., 2016; Marete A. G. et al., 2018; Bouwman et al., 2018; Tribout et al., 2020). Within this region, for example, we find the gene *XKR4* associated with subcutaneous rump fat thickness and growth (Neto et al., 2012; Magalhã et al., 2016; An et al., 2019; Smith et al., 2019). Lastly, the region on BTA 18 was reported as a candidate region for tuberculosis resistance (Ring et al., 2019), milk (Cole et al., 2011; Ibeagha-Awemu et al., 2016; Benedet et al., 2019), and fertility traits (Cole et al., 2011; Gaddis et al., 2016) including spermatogenesis-associated proteins 33 and 2L (*SPATA33* and *SPATA2L*) in the extended regions of 250 kb.

Regions detected in DSN, but not in Holstein, were very often associated with meat and carcass. Among the RoH islands detected in each breed 76.2% in DSN were associated with meat, but only 47.4% in Holstein. For milk, 95.2% of RoH islands in DSN and 78.9% of RoHs islands in Holstein were associated with milk production. Interestingly, the RoH region on BTA 28, which was found in DSN only, previously has been associated with milk production traits in a genome-wide association study in DSN (Korkuć et al., 2021).

Although differences were seen in signatures of selection, many signatures are shared between DSN and Holstein. This is consistent with the genetic relatedness and shared history of DSN and Holstein and the fact that both breeds had been selected for milk yield.

Considering all the results from signatures of selection, traits related to meat and milk showed the largest differences between DSN and Holstein, due to different selection goals. Genes affecting fertility, exterior, production, and health traits were also very frequent. Those findings are consistent with the characteristics of the breed-type and purposes described by the breeding organizations.

5 Conclusion

Despite the small population size of 2,500 animals, the DSN breed does not show any signs of loss of diversity or increased inbreeding compared to other taurine breeds. On the contrary, the inbreeding degree in DSN is even lower and the diversity higher than in Holstein. This is a remarkable result of the breeding strategy used for the maintenance of DSN as a genetic resource and shows the potential of maintaining small local populations while keeping diversity and controlling inbreeding. Our study provides the background for cattle breeds that are closely related to DSN and could, therefore, serve as an external gene pool to keep or even increase the diversity in DSN. Our analyses also provide evidence for high genomic diversity in breeds such as Yakut, Charolais, Kholmogory, and Modern Danish Red, while inbreeding was high in Jersey, Wagyu, Hereford, and Shorthorn, pointing to extra care needed for those breeds.

Moreover, specific genomic regions and positional candidate genes seem to be partially responsible for the DSN-specific characteristics. These include candidate genes previously identified in association studies with DSN, such as one region detected in DSN for endoparasite infection resistance, an important trait for pasture systems. In addition, these regions point to genes associated with traits that have not been studied yet in DSN, but in other breeds or species. Such regions are likely of particular interest for the conservation of DSN and the maintenance of its specific characteristics. Further studies are needed in order to elucidate the function of those regions and underlying causal sequence variants. Besides investigating milk and beef production, the study of new traits for disease resistance and resilience, such as heat stress, methane emissions or feed uptake capacity can further improve our understanding of the importance of DSN as a small local breed and as a genetic resource that contributes to conserve the whole genomic diversity of the species.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ebi.ac.uk/ena/browser/view/PRJEB45822. Data from other breeds was provided within the frame of the 1000 Bull Genomes Project Consortium (Run 9).

Ethics statement

Ethical review and approval was not required for the animal study because samples were previously collected based on routine procedures on these farm animals. Ear tags were taken as part of the required registration procedure, blood samples were taken by a trained veterinarian to perform standard health recording.

Written informed consent was obtained from the owners for the participation of their animals in this study.

Author contributions

GB, GN, PK, and DA designed the study. GN performed all analyses under PK, DA, and GB supervision. MW helped with the XP-EHH analysis. GN, PK, DA, and GB drafted the manuscript. All authors read and approved the final manuscript.

Funding

This work was financially supported by the German Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany, granted by the Federal Office for Agriculture and Food (BLE), grant number FKZ2818BM090 and 091. The article processing charge was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—491192747 and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

Acknowledgments

We acknowledge the 1000 Bull Genomes Consortium for providing the data, the Rinderproduktion Berlin-Brandenburg GmbH and DSN farms for supporting this project with data, tissue collection and their expertise in animal selection, and Sèyi Fridaïus Ulrich Vanvanhossou for help with the XP-EHH calculations in R.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.993959/full#supplementary-material

References

Afrache, H., Gouret, P., Ainouche, S., Pontarotti, P., and Olive, D. (2012). The butyrophilin (BTN) gene family: From milk fat to the regulation of the immune response. *Immunogenetics* 64, 781–794. doi:10.1007/s00251-012-0619-z

Akanno, E. C., Chen, L., Abo-Ismail, M. K., Crowley, J. J., Wang, Z., Li, C., et al. (2018). Genome-wide association scan for heterotic quantitative trait loci in multi-breed and crossbred beef cattle. *Genet. Sel. Evol.* 50, 48. doi:10.1186/s12711-018-0405-y

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

Ali, A. A., Khatkar, M. S., Kadarmideen, H. N., and Thomson, P. C. (2015). Additive and epistatic genome-wide association for growth and ultrasound scan measures of carcass-related traits in Brahman cattle. *J. Anim. Breed. Genet.* 132, 187–197. doi:10.1111/jbg.12147

Almodóvar-Payá, A., Villarreal-Salazar, M., Luna, N., Real-Martínez, A., Andreu, A. L., et al. (2020). Preclinical research in glycogen storage diseases: A comprehensive review of current animal models. *Int. J. Mol. Sci.* 9621 (21), 9621. doi:10.3390/ijms21249621

An, B., Xia, J., Chang, T., Wang, X., Xu, L., Zhang, L., et al. (2019). Genome-wide association study reveals candidate genes associated with body measurement traits in Chinese Wagyu beef cattle. *Anim. Genet.* 50, 386–390. doi:10.1111/age.12805

Bellamy, W., Takase, M., Yamauchi, K., Wakabayashi, H., Kawase, K., and Tomita, M. (1992). Identification of the bactericidal domain of lactoferrin. *Biochim. Biophys. Acta* 1121, 130–136. doi:10.1016/0167-4838(92)90346-f

Benedet, A., Ho, P. N., Xiang, R., Bolormaa, S., De Marchi, M., Goddard, M. E., et al. (2019). The use of mid-infrared spectra to map genes affecting milk composition. *J. Dairy Sci.* 102, 7189–7203. doi:10.3168/jds.2018-15890

Bennewitz, J., Reinsch, N., Grohs, C., Levéziel, H., Malafosse, A., Thomsen, H., et al. (2003). Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and increasing experimental power in dairy cattle. *Genet. Sel. Evol.* 353 (35), 319–338. doi:10.1186/1297-9686-35-3-319

Berko, E. R., Cho, M. T., Eng, C., Shao, Y., Sweetser, D. A., Waxler, J., et al. (2017). De novo missense variants in HECW2 are associated with neurodevelopmental delay and hypotonia. *J. Med. Genet.* 54, 84–86. doi:10.1136/jmedgenet-2016-103943

Bhati, M., Kadri, N. K., Crysnanto, D., and Pausch, H. (2020). Assessing genomic diversity and signatures of selection in Original Braunvieh cattle using whole-genome sequencing data. *BMC Genomics* 21, 27. doi:10.1186/s12864-020-6446-y

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Res.* 23, 1514–1521. doi:10.1101/gr.154831.113

Biedermann, G. (2003). Zuchtplanung für die Erhaltung des Alten Schwarzbunten Niederungsrindes. Available at: https://service.ble.de/ptdb/index2.php?detail_id=84928&site_key=141.

Biscarini, F., Nicolazzi, E., Alessandra, S., Boettcher, P., and Gandini, G. (2015). Challenges and opportunities in genetic improvement of local livestock breeds. *Front. Genet.* 5, 33–16. doi:10.3389/fgene.2015.00033

Boettcher, P. J., Hoffmann, I., Baumung, R., Drucker, A. G., McManus, C., Berg, P., et al. (2014). Genetic resources and genomics for adaptation of livestock to climate change. *Front. Genet.* 5, 461. doi:10.3389/fgene.2014.00461

Bohlouli, M., Halli, K., Yin, T., Gengler, N., and König, S. (2022). Genome-wide associations for heat stress response suggest potential candidate genes underlying milk fatty acid composition in dairy cattle. *J. Dairy Sci.* 105, 3323–3340. doi:10.3168/jds.2021-21152

Bolormaa, S., Neto, P., Zhang, Y. D., Bunch, R. J., Harrison, B. E., Goddard, M. E., et al. (2011). A genome-wide association study of meat and carcass traits in Australian cattle. *J. Anim. Sci.* 89, 2297–2309. doi:10.2527/jas.2010-3138

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

Brade, W. (2011). Die Milchrinderzüchtung in der ehemaligen DDR eine retrospektive Bewertung. *Berichte uber Landwirtsch.* 89, 267.

Brade, W., and Brade, E. (2013). "Breeding history of German Holstein cattle," in *Berichte über Landwirtschaft* (Hannover). https://www.cabdirect.org/cabdirect/abstract/20133389715.

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi:10.1016/j.ajhg.2018.07.015

BRS (2021). Black and White dual purpose. Available at: https://www.rind-schwein.de/brs-cattle/black-and-white-dual-purpose-en.html.

Butty, A. M., Frischknecht, M., Gredler, B., Neuenschwander, S., Moll, J., Bieber, A., et al. (2017). Genetic and genomic analysis of hyperthelia in Brown Swiss cattle. *J. Dairy Sci.* 100, 402–411. doi:10.3168/jds.2016-11420

Cai, Z., Dusza, M., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2020). Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. *Genet. Sel. Evol.* 52, 19. doi:10.1186/s12711-020-00538-6

Cecchinato, A., Ribeca, C., Chessa, S., Cipolat-Gotet, C., Maretto, F., Casellas, J., et al. (2014). Candidate gene association analysis for milk yield, composition, urea nitrogen and somatic cell scores in Brown Swiss cows. *Animal* 8, 1062–1070. doi:10.1017/S1751731114001098

Choi, K. S., Choi, H. J., Lee, J. K., Im, S., Zhang, H., Jeong, Y., et al. (2016). The endothelial E3 ligase HECW2 promotes endothelial cell junctions by increasing AMOTL1 protein stability via K63-linked ubiquitination. *Cell. Signal.* 28, 1642–1651. doi:10.1016/j.cellsig.2016.07.015

Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., et al. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics* 12, 408. doi:10.1186/1471-2164-12-408

Connor, E. E., Zhou, Y., and Liu, G. E. (2018). The essence of appetite: Does olfactory receptor variation play a role? *J. Anim. Sci.* 96, 1551–1558. doi:10.1093/jas/sky068

Crispim, A. C., Kelly, M. J., Guimarães, S. E. F., Silva, E., Fortes, M. R. S., Wenceslau, R. R., et al. (2015). Multi-trait GWAS and new candidate genes annotation for growth curve parameters in brahman cattle. *PLoS One* 10, e0139906. doi:10.1371/journal.pone.0139906

da Silva, Sinani, D., and Jones, C. (2012). ICP27 protein encoded by bovine herpesvirus type 1 (bICP27) interferes with promoter activity of the bovine genes encoding beta interferon 1 (IFN-β1) and IFN-β3. *Virus Res.* 169, 162–168. doi:10.1016/j.virusres.2012.07.023

Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M., and Robinson, J. A. B. (2008). A genome scan to detect quantitative trait loci for economically important traits in holstein cattle using two methods and a dense single nucleotide polymorphism map. *J. Dairy Sci.* 91, 3225–3236. doi:10.3168/jds.2007-0333

D'Amours, O., Bordeleau, L. J., Frenette, G., Blondin, P., Leclerc, P., and Sullivan, R. (2012). Binder of sperm 1 and epididymal sperm binding protein 1 are associated with different bull sperm subpopulations. *Reproduction* 143, 759–771. doi:10.1530/REP-11-0392

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

de Simoni Gouveia, J. J., Paiva, S. R., McManus, C. M., Caetano, A. R., Kijas, J. W., Facó, O., et al. (2017). Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds. *Livest. Sci.* 197, 36–45. doi:10.1016/j.livsci.2017.01.006

Dixit, S. P., Singh, S., Ganguly, I., Bhatia, A. K., Sharma, A., Kumar, N. A., et al. (2020). Genome-wide runs of homozygosity revealed selection signatures in *Bos indicus*. *Front. Genet.* 11, 92. doi:10.3389/fgene.2020.00092

Dmitriev, N., and Ernst, L. (1987). Animal genetic resources of the USSR. *Anim. Prod. Heal. Pap. Publ. by FAO* 65, 18

Dutta, P., Talenti, A., Young, R., Jayaraman, S., Callaby, R., Jadhav, S. K., et al. (2020). Whole genome analysis of water buffalo and global cattle breeds highlights convergent signatures of domestication. *Nat. Commun.* 111 (11), 4739–4813. doi:10.1038/s41467-020-18550-1

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Mol. Ecol.* 14, 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

FAO (2007). *Global plan of action for animal genetic resources and the interlaken declaration*. Rome, Italy: FAO. Available at: https://www.fao.org/publications/card/en/c/dafd2e54-21d2-51cc-a79d-457fb447a11b/

FAO (2011). Molecular genetic characterization of animal genetic resources." in *Commission on Genetic resources for Food and Agriculture, Food and Agriculture Organization of the United Nations*.

FAO (2021). Risk status of livestock breeds. Available at: https://www.fao.org/sustainable-development-goals/indicators/252/en/.

Felius, M., Beerling, M. L., Buchanan, D. S., Theunissen, B., Koolmees, P. A., and Lenstra, J. A. (2014). On the history of cattle genetic resources. *Diversity* 6, 705–750. doi:10.3390/d6040705

Felius, M., Koolmees, P. A., Theunissen, B., Lenstra, J. A., Baumung, R., Manatrinon, S., et al. (2011). On the breeds of cattle—historic and current classifications. *Divers. (Basel).* 3, 660–692. doi:10.3390/d3040660

Fitak, R. R. (2021). OptM: Estimating the optimal number of migration edges on population trees using treemix. *Biol. Methods Protoc.* 6, bpab017. doi:10.1093/biomethods/bpab017

Forutan, M., Ansari Mahyari, S., Baes, C., Melzer, N., Schenkel, F. S., and Sargolzaei, M. (2018). Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. *BMC Genomics* 19, 98. doi:10.1186/s12864-018-4453-z

Frankham, R. (1997). Do island populations have less genetic variation than mainland populations? *Heredity* 783 (78), 311–327. doi:10.1038/hdy.1997.46

Freua, M. C., Santana, M. H. A., Ventura, R. V., and Ferraz, J. B. S. (2016). Parameters of a dynamic mechanistic model of cattle growth retain enough biological interpretation for genotype-to-phenotype mapping. *Genet. Mol. Res.* 15, 15048931. doi:10.4238/gmr15048931

Gaddis, 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

Gandini, G., Maltecca, C., Pizzi, F., Bagnato, A., and Rizzi, R. (2007). Comparing local and commercial breeds on functional traits and profitability: The case of reggiana dairy cattle. *J. Dairy Sci.* 90, 2004–2011. doi:10.3168/jds.2006-204

Gao, Y., Bai, C., Zheng, D., Li, C., Zhang, W., Li, M., et al. (2016). Combination of melatonin and Wnt-4 promotes neural cell differentiation in bovine amniotic epithelial cells and recovery from spinal cord injury. *J. Pineal Res.* 60, 303–312. doi:10.1111/jpi.12311

Gautier, M., Klassmann, A., and Vitalis, R. (2017). Rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* 17, 78–90. doi:10.1111/1755-0998.12634

GEH e.V (2020). *Die Rote Liste im Überblick*. Kassel, Germany: Soc. Conserv. Old Endanger. Livest. Breeds. https://www.g-e-h.de/index.php/rote-liste-menu/rote-liste.

Gorssen, W., Meyermans, R., Janssens, S., and Buys, N. (2021). A publicly available repository of ROH islands reveals signatures of selection in different livestock and pet species. *Genet. Sel. Evol.* 531 (53), 2–10. doi:10.1186/s12711-020-00599-7

Grati, M., Chakchouk, I., Ma, Q., Bensaid, M., Desmidt, A., Turki, N., et al. (2015). A missense mutation in DCDC2 causes human recessive deafness DFNB66, likely by interfering with sensory hair cell and supporting cell cilia length regulation. *Hum. Mol. Genet.* 24, 2482–2491. doi:10.1093/hmg/ddv009

Grigoletto, L., Ferraz, J. B. S., Oliveira, H. R., Eler, J. P., Bussiman, F. O., Abreu Silva, B. C., et al. (2020a). Genetic architecture of carcass and meat quality traits in Montana Tropical® composite beef cattle. *Front. Genet.* 11, 123. doi:10.3389/fgene.2020.00123

Grigoletto, L., Santana, M. H. A., Bressan, F. F., Eler, J. P., Nogueira, M. F. G., Kadarmideen, H. N., et al. (2020b). Genetic parameters and genome-wide association studies for anti-müllerian hormone levels and antral follicle populations measured after estrus synchronization in Nellore cattle. *Animals.* 10, 1185–1215. doi:10.3390/ani10071185

Gutiérrez-Reinoso, M. A., Aponte, P. M., and García-Herreros, M. (2022). A review of inbreeding depression in dairy cattle: Current status, emerging control strategies, and future prospects. *J. Dairy Res.* 89, 3–12. doi:10.1017/S0022029922000188

Hamidi Hay, E., and Roberts, A. (2017). Genomic prediction and genome-wide association analysis of female longevity in a composite beef cattle breed. *J. Anim. Sci.* 95, 1467–1471. doi:10.2527/jas.2016.1355

Hayes, B. J., and Daetwyler, H. D. (2019). 1000 bull genomes project to map simple and complex genetic traits in cattle: Applications and outcomes. *Annu. Rev. Anim. Biosci.* 7, 89–102. doi:10.1146/annurev-animal-020518-115024

Hiemstra, S. J., Haas, Y., and Gandini, G. (2010). *Local cattle breeds in Europe*. 1st ed. Wageningen, The Netherlands: Wageningen Academic Publishers. doi:10.3920/978-90-8686-697-7

Hoffmann, I. (2011). Livestock biodiversity and sustainability. *Livest. Sci.* 139, 69–79. doi:10.1016/j.livsci.2011.03.016

Hu, Z.-L., Park, C. A., and Reecy, J. M. (2022). Bringing the animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services. *Nucleic Acids Res.* 50, D956–D961. doi:10.1093/nar/gkab1116

Huang, Y., Chen, B., Ye, M., Liang, P., Zhangfang, Y., Huang, J., et al. (2016). Ccndbp1 is a new positive regulator of skeletal myogenesis. *J. Cell. Sci.* 129, 2767–2777. doi:10.1242/jcs.184234

Hudson, R. R., Slatkint, M., and Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* 132, 583–589. doi:10.1093/genetics/132.2.583

Hulsegge, I., Oldenbroek, K., Bouwman, A., Veerkamp, R., and Windig, J. (2022). Selection and drift: A comparison between historic and recent Dutch friesian cattle and recent holstein friesian using WGS data. *Animals.* 329 (12), 329. doi:10.3390/ani12030329

Huson, H. J., Sonstegard, T. S., Godfrey, J., Hambrook, D., Wolfe, C., Wiggans, G., et al. (2020). A genetic investigation of island Jersey cattle, the foundation of the Jersey breed: Comparing population structure and selection to Guernsey, holstein, and United States Jersey cattle. *Front. Genet.* 11, 366. doi:10.3389/fgene.2020.00366

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

Jaeger, M., Scheper, C., König, S., and Brügemann, K. (2018). Inbreeding and genetic relationships of the endangered dual-purpose black and white cattle breed (DSN) based on own genetic breed percentage calculations. *Züchtungskd.* 90, 262.

Jiang, J., Ma, L., Prakapenka, D., VanRaden, P. M., Cole, J. B., and Da, Y. (2019). A large-scale genome-wide association study in U.S. Holstein cattle. *Front. Genet.* 10, 412. doi:10.3389/fgene.2019.00412

Jumeau, F., Chalmel, F., Fernandez-Gomez, F. J., Carpentier, C., Obriot, H., Tardivel, M., et al. (2017). Defining the human sperm microtubulome: An integrated genomics approach. *Biol. Reprod.* 96, 93–106. doi:10.1095/biolreprod.116.143479

Kardos, M., Armstrong, E. E., Fitzpatrick, S. W., Hauser, S., Hedrick, P. W., Miller, J. M., et al. (2021). The crucial role of genome-wide genetic variation in conservation. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2104642118. doi:10.1073/pnas.2104642118

Kelleher, M. M., Berry, D. P., Kearney, J. F., McParland, S., Buckley, F., and Purfield, D. C. (2017). Inference of population structure of purebred dairy and beef cattle using high-density genotype data. *Animal* 11, 15–23. doi:10.1017/S1751731116001099

Kim, E. S., Cole, J. B., Huson, H., Wiggans, G. R., Tassel, Van, Crooker, B. A., et al. (2013). Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. *PLoS One* 8, 813. doi:10.1371/journal.pone.0080813

Kim, E. T., Joo, S. S., Kim, D. H., Gu, B. H., Park, D. S., Atikur, R. M., et al. (2021). Common and differential dynamics of the function of peripheral blood mononuclear cells between holstein and Jersey cows in heat-stress environment. *Animals.* 11, 19–20. doi:10.3390/ani11010019

Kim, H. J., Sharma, A., Lee, S. H., Lee, D. H., Lim, D. J., Cho, Y. M., et al. (2017). Genetic association of PLAG1, SCD, CYP7B1 and FASN SNPs and their effects on carcass weight, intramuscular fat and fatty acid composition in Hanwoo steers (Korean cattle). *Anim. Genet.* 48, 251–252. doi:10.1111/age.12523

Klisch, K., Sousa, De, Beckers, J. F., Leiser, R., and Pich, A. (2005). Pregnancy associated glycoprotein-1, -6, -7, and -17 are major products of bovine binucleate trophoblast giant cells at midpregnancy. *Mol. Reprod. Dev.* 71, 453–460. doi:10.1002/mrd.20296

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

Korkuć, P., Arends, D., May, K., König, S., and Brockmann, G. A. (2021). Genomic loci affecting milk production in German black pied cattle (DSN). *Front. Genet.* 12, 640039. doi:10.3389/fgene.2021.640039

Kristensen, T. N., Hoffmann, A. A., Pertoldi, C., and Stronen, A. V. (2015). What can livestock breeders learn from conservation genetics and vice versa? *Front. Genet.* 6, 38. doi:10.3389/fgene.2015.00038

Kunzmann, P., and Hamacher, K. (2018). Biotite: A unifying open source computational biology framework in Python. *BMC Bioinforma. 2018* 191 (19), 1–8. doi:10.1186/S12859-018-2367-Z

Lacroix-Pépin, N., Danyod, G., Krishnaswamy, N., Mondal, S., Rong, P. M., Chapdelaine, P., et al. (2011). The multidrug resistance-associated protein 4 (MRP4) appears as a functional carrier of prostaglandins regulated by oxytocin in the bovine endometrium. *Endocrinology* 152, 4993–5004. doi:10.1210/en.2011-1406

Lee, Y. C., Chang, C. J., Bali, D., Chen, Y. T., and Yan, Y. T. (2011). Glycogen-branching enzyme deficiency leads to abnormal cardiac development: Novel insights into glycogen storage disease IV. *Hum. Mol. Genet.* 20, 455–465. doi:10.1093/hmg/ddq492

Letunic, I., and Bork, P. (2019). *Interactive tree of life (iTOL) v4: Recent updates and new developments*. Web Serv. issue Publ. doi:10.1093/nar/gkz239

Li, C. C., and Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* 5, 107.

López-Gatius, F., Hunter, R. H. F., Garbayo, J. M., Santolaria, P., Yániz, J., Serrano, B., et al. (2007). Plasma concentrations of pregnancy-associated glycoprotein-1 (PAG-1) in high producing dairy cows suffering early fetal loss during the warm season. *Theriogenology* 67, 1324–1330. doi:10.1016/j.theriogenology.2007.02.004

Lorenz, M. D., Coates, J. R., and Kent, M. (2011). Tetraparesis, hemiparesis, and ataxia. *Handb. Vet. Neurol.*, 162–249. doi:10.1016/B978-1-4377-0651-2.10007-4

Lu, D., Miller, S., Sargolzaei, M., Kelly, M., Vander Voort, G., Caldwell, T., et al. (2013). Genome-wide association analyses for growth and feed efficiency traits in beef cattle. *J. Anim. Sci.* 91, 3612–3633. doi:10.2527/jas.2012-5716

Magalhã, A. F. B., De Camargo, G. M. F., Fernandes, J., Gordo, D. G. M., Tonussi, R. L., Costa, R. B., et al. (2016). Genome-wide association study of meat quality traits in Nellore cattle. *PLoS One* 11, e0157845. doi:10.1371/journal.pone.0157845

Makanjuola, B. O., Maltecca, C., Miglior, F., Schenkel, F. S., and Baes, C. F. (2020). Effect of recent and ancient inbreeding on production and fertility traits in Canadian Holsteins. *BMC Genomics* 21, 605–615. doi:10.1186/s12864-020-07031-w

Marete, A. G., Guldbrandtsen, B., Lund, M. S., Fritz, S., Sahana, G., and Boichard, D. (2018a). A meta-analysis including pre-selected sequence variants associated with seven traits in three French dairy cattle populations. *Front. Genet.* 9, 522. doi:10.3389/fgene.2018.00522

Marete, A., Sahana, G., Fritz, S., Lefebvre, R., Barbat, A., Lund, M. S., et al. (2018b). Genome-wide association study for milking speed in French Holstein cows. *J. Dairy Sci.* 101, 6205–6219. doi:10.3168/jds.2017-14067

Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C., Ajmone-Marsan, P., Valentini, A., et al. (2015). Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Anim. Genet.* 46, 110–121. doi:10.1111/age.12259

Martínez, R., Gómez, Y., and Martínez-Roch, J. F. (2014). Genome-wide association study on growth traits in Colombian creole breeds and crossbreeds with Zebu cattle. *Genet. Mol. Res.* 13, 6420–6432. doi:10.4238/2014.August.25.5

Mastrangelo, S., Tolone, M., Di Gerlando, R., Fontanesi, L., Sardina, M. T., and Portolano, B. (2016). Genomic inbreeding estimation in small populations: Evaluation of runs of homozygosity in three local dairy cattle breeds. *Animal* 10, 746–754. doi:10.1017/S1751731115002943

May, K., Scheper, C., Brügemann, K., Yin, T., Strube, C., Korkuć, P., et al. (2019). Genome-wide associations and functional gene analyses for endoparasite resistance in an endangered population of native German Black Pied cattle. *BMC Genomics* 20, 277. doi:10.1186/s12864-019-5659-4

McClure, M. C., Morsci, N. S., Schnabel, R. D., Kim, J. W., Yao, P., Rolf, M. M., et al. (2010). A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. *Anim. Genet.* 41, 597–607. doi:10.1111/j.1365-2052.2010.02063.x

McClure, M. C., Ramey, H. R., Rolf, M. M., McKay, S. D., Decker, J. E., Chapple, R. H., et al. (2012). Genome-wide association analysis for quantitative trait lociinfluencing Warner–Bratzler shear force in five taurine cattlebreeds. *Anim. Genet.* 43, 662–673. doi:10.1111/j.1365-2052.2012.02323.x

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

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

Meier, S., Arends, D., Korkuć, P., Neumann, G. B., and Brockmann, G. A. (2020). A genome-wide association study for clinical mastitis in the dual-purpose German Black Pied cattle breed. *J. Dairy Sci.* 103, 10289–10298. doi:10.3168/jds.2020-18209

Meredith, B. K., Kearney, F. J., Finlay, E. K., Bradley, D. G., Fahey, A. G., Berry, D. P., et al. (2012). Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. *BMC Genet.* 13, 21. doi:10.1186/1471-2156-13-21

Mészáros, G., Boison, S. A., O'Brien, Pérez, Ferencakovic, 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

Miles, A., Ralph, P., Harding, N., Pisupati, R., et al. (2020). *cggh/scikit-allel*. doi:10.5281/ZENODO.3935797

Naderi, S., Moradi, M. H., Farhadian, M., Yin, T., Jaeger, M., Scheper, C., et al. (2020). Assessing selection signatures within and between selected lines of dual-purpose black and white and German Holstein cattle. *Anim. Genet.* 51, 391–408. doi:10.1111/age.12925

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: A hidden markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32, 1749–1751. doi:10.1093/bioinformatics/btw044

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273. doi:10.1073/pnas.76.10.5269

Neto, P., Bunch, R. J., Harrison, B. E., and Barendse, W. (2012). Variation in the XKR4 gene was significantly associated with subcutaneous rump fat thickness in indicine and composite cattle. *Anim. Genet.* 43, 785–789. doi:10.1111/j.1365-2052.2012.02330.x

Neumann, G. B., Korkuć, P., Arends, D., Wolf, M. J., May, K., Reißmann, M., et al. (2021). Design and performance of a bovine 200 k SNP chip developed for endangered German Black Pied cattle (DSN). *BMC Genomics* 221 (22), 905–913. doi:10.1186/s12864-021-08237-2

Olivieri, B. F., Mercadante, M. E. Z., Cyrillo, J. N. D. S. G., Branco, R. H., Bonilha, S. F. M., De 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

Olsen, H. G., Gomez-Raya, L., Våge, D. I., Olsaker, I., Klungland, H., Svendsen, M., et al. (2002). A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. *J. Dairy Sci.* 85, 3124–3130. doi:10.3168/jds.S0022-0302(02)74400-7

Pausch, H., Emmerling, R., Schwarzenbacher, H., and Fries, R. (2016). A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. *Genet. Sel. Evol.* 48, 14. doi:10.1186/s12711-016-0190-4

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

Pierce, A., Colavizza, D., Benaissa, M., Maes, P., Tartar, A., Montreuil, J., et al. (1991). Molecular cloning and sequence analysis of bovine lactotransferrin. *Eur. J. Biochem.* 196, 177–184. doi:10.1111/j.1432-1033.1991.tb15801.x

Pryce, J. E., Hayes, B. J., Bolormaa, S., and Goddard, M. E. (2011). Polymorphic regions affecting human height also control stature in cattle. *Genetics* 187, 981–984. doi:10.1534/genetics.110.123943

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

Rejman, J. J., Hegarty, H. M., and Hurley, W. L. (1989). Purification and characterization of bovine lactoferrin from secretions of the involuting mammary gland: Identification of multiple molecular weight forms. *Comp. Biochem. Physiol. B* 93, 929–934. doi:10.1016/0305-0491(89)90068-0

Rezende, F. M., Rodriguez, E., Leal-Gutiérrez, J. D., Elzo, M. A., Johnson, D. D., Carr, C., et al. (2021). Genomic approaches reveal pleiotropic effects in crossbred beef cattle. *Front. Genet.* 12. doi:10.3389/fgene.2021.627055

Ring, S. C., Purfield, D. C., Good, M., Breslin, P., Ryan, E., Blom, A., et al. (2019). Variance components for bovine tuberculosis infection and multi-breed genome-wide association analysis using imputed whole genome sequence data. *PLoS One* 14, e0212067. doi:10.1371/journal.pone.0212067

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., et al. (2020). *De novo* assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* 9, 1–9. doi:10.1093/gigascience/giaa021

Rotin, D., and Kumar, S. (2009). Physiological functions of the HECT family of ubiquitin ligases. *Nat. Rev. Mol. Cell. Biol.* 106 (10), 398–409. doi:10.1038/nrm2690

Rupp, R., and Boichard, D. (2003). Genetics of resistance to mastitis in dairy cattle. *Vet. Res.* 34, 671–688. doi:10.1051/vetres:2003020

Rutten, M. J. M., Bouwman, A. C., Sprong, R. C., van Arendonk, J. A. M., and Visker, M. H. P. W. (2013). Genetic variation in vitamin B-12 content of bovine milk and its association with SNP along the bovine genome. *PLoS One* 8, e62382. doi:10.1371/journal.pone.0062382

Saatchi, M., Schnabel, R. D., Taylor, J. F., and Garrick, D. J. (2014). Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genomics* 15, 442. doi:10.1186/1471-2164-15-442

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* 4497164 (449), 913–918. doi:10.1038/nature06250

Sanchez, M. P., Govignon-Gion, A., Croiseau, P., Fritz, S., Hozé, C., Miranda, G., et al. (2017). Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genet. Sel. Evol.* 49, 68. doi:10.1186/s12711-017-0344-z

Sanchez, M. P., Ramayo-Caldas, Y., Wolf, V., Laithier, C., El Jabri, M., Michenet, A., et al. (2019). Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. *Genet. Sel. Evol.* 51, 34. doi:10.1186/s12711-019-0473-7

Schmidtmann, C., Schönherz, A., Guldbrandtsen, B., Marjanovic, J., Calus, M., Hinrichs, D., et al. (2021). Assessing the genetic background and genomic relatedness of red cattle populations originating from Northern Europe. *Genet. Sel. Evol.* 531 (53), 23–18. doi:10.1186/s12711-021-00613-6

Seabold, S., and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python." in Proceedings of the 9th Python in Science Conference

Sharma, A., Dang, C. G., Kim, K. S., Kim, J. J., Lee, H. K., Kim, H. C., et al. (2014). Validation of genetic polymorphisms on BTA14 associated with carcass trait in a commercial Hanwoo population. *Anim. Genet.* 45, 863–867. doi:10.1111/age.12204

Shiina, T., Blancher, A., Inoko, H., and Kulski, J. K. (2017). Comparative genomics of the human, macaque and mouse major histocompatibility complex. *Immunology* 150, 127–138. doi:10.1111/imm.12624

Smith, J. L., Wilson, M. L., Nilson, S. M., Rowan, T. N., Oldeschulte, D. L., Schnabel, R. D., et al. (2019). Genome-wide association and genotype by environment interactions for growth traits in U.S. Gelbvieh cattle. *BMC Genomics* 20, 926–1013. doi:10.1186/s12864-019-6231-y

Snelling, W. M., Allan, M. F., Keele, J. W., Kuehn, L. A., McDaneld, T., Smith, T. P. L., et al. (2010). Genome-wide association study of growth in crossbred beef cattle. *J. Anim. Sci.* 88, 837–848. doi:10.2527/jas.2009-2257

Soares, R. A. N., Vargas, G., Duffield, T., Schenkel, F., and Squires, E. J. (2021). Genome-wide association study and functional analyses for clinical and subclinical ketosis in Holstein cattle. *J. Dairy Sci.* 104, 10076–10089. doi:10.3168/jds.2020-20101

Song, Y., Xu, L., Chen, Y., Zhang, L., Gao, H., Zhu, B., et al. (2016). Genome-wide association study reveals the PLAG1 gene for knuckle, biceps and shank weight in simmental beef cattle. *PLoS One* 11, e0168316. doi:10.1371/journal.pone.0168316

Srikanth, K., Lee, S. H., Chung, K. Y., Park, J. E., Jang, G. W., Park, M. R., et al. (2020). A gene-set enrichment and protein–protein interaction network-based GWAS with regulatory SNPs identifies candidate genes and pathways associated with carcass traits in hanwoo cattle. *Genes.* 11, 316. doi:10.3390/genes11030316

Stier, K. (20212022). *Die Rote Liste der gefährdeten Nutztierrassen der GEH*. GEH. Available at: https://www.g-e-h.de/index.php/nachrichten/330-neueroteliste.

TGRDEU (2021). Rind: Deutsches Schwarzbuntes Niederungsrind. Available at: https://tgrdeu.genres.de/nutztiere/suche-nutztiere/genetik-detaildarstellung/?tx_sttgrdeu_nutztier%5Baction%5D=genetikDetail&tx_sttgrdeu_nutztier%5Bcontroller%5D=Nutztier&tx_sttgrdeu_nutztier%5Bg_id%5D=654&cHash=961248c5168bd52ab4538b0d7506f9e8.

Tribout, T., Croiseau, P., Lefebvre, R., Barbat, A., Boussaha, M., Fritz, S., et al. (2020). Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. *Genet. Sel. Evol.* 52, 55. doi:10.1186/s12711-020-00575-1

Tríbulo, P., Siqueira, L. G. B., Oliveira, L. J., Scheffler, T., and Hansen, P. J. (2018). Identification of potential embryokines in the bovine reproductive tract. *J. Dairy Sci.* 101, 690–704. doi:10.3168/jds.2017-13221

Van Den Berg, I., Xiang, R., Jenko, J., Pausch, H., Boussaha, M., Schrooten, C., et al. (2020). Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94, 321 cattle from eight cattle breeds. *Genet. Sel. Evol.* 52, 37. doi:10.1186/s12711-020-00556-4

Vanvanhossou, S. F. U., Scheper, C., Dossa, L. H., Yin, T., Brügemann, K., and König, S. (2020). A multi-breed GWAS for morphometric traits in four Beninese indigenous cattle breeds reveals loci associated with conformation, carcass and adaptive traits. *BMC Genomics* 211 (21), 783–816. doi:10.1186/s12864-020-07170-0

Walter, S., Atzmon, G., Demerath, E. W., Garcia, M. E., Kaplan, R. C., Kumari, M., et al. (2011). A genome-wide association study of aging. *Neurobiol. Aging* 32, e15–e28. doi:10.1016/j.neurobiolaging.2011.05.026

Wang, S., Raza, S. H. A., Zhang, K., Mei, C., Alamoudi, M. O., Aloufi, B. H., et al. (2022). Selection signatures of Qinchuan cattle based on whole-genome sequences. *Anim. Biotechnol.* 1, 3252. doi:10.1080/10495398.2022.2033252

Wang, Y., Zhang, F., Mukiibi, R., Chen, L., Vinsky, M., Plastow, G., et al. (2020). Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: II: Carcass merit traits. *BMC Genomics* 21, 6273. doi:10.1186/S12864-019-6273-1

Ward, T. L., Valberg, S. J., Adelson, D. L., Abbey, C. A., Binns, M. M., and Mickelson, J. R. (2004). Glycogen branching enzyme (GBE1) mutation causing equine glycogen storage disease IV. *Mamm. Genome* 15, 570–577. doi:10.1007/s00335-004-2369-1

Wei, Q., Zhong, L., Zhang, S., Mu, H., Xiang, J., Yue, L., et al. (2017). Bovine lineage specification revealed by single-cell gene expression analysis from zygote to blastocyst. *Biol. Reprod.* 97, 5–17. doi:10.1093/biolre/iox071

Weldenegodguad, M., Popov, R., Pokharel, K., Ammosov, I., Ming, Y., Ivanova, Z., et al. (2019). Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. *Front. Genet.* 10, 728. doi:10.3389/fgene.2018.00728

Wolf, M. J., Yin, T., Neumann, G. B., Korkuć, P., Brockmann, G. A., König, S., et al. (2021). Genome-wide association study using whole-genome sequence data for fertility, health indicator, and endoparasite infection traits in German black pied cattle. *Genes*. 1163 (12), 1163. doi:10.3390/genes12081163

Wu, S., Mipam, T. D., Xu, C., Zhao, W., Shah, M. A., Yi, C., et al. (2020). Testis transcriptome profiling identified genes involved in spermatogenic arrest of cattleyak. *PLoS One* 15, e0229503. doi:10.1371/journal.pone.0229503

Wu, X., Guldbrandtsen, B., Lund, M. S., and Sahana, G. (2016). Association analysis for feet and legs disorders with whole-genome sequence variants in 3 dairy cattle breeds. *J. Dairy Sci.* 99, 7221–7231. doi:10.3168/jds.2015-10705

Xia, X., Zhang, S., Zhang, H., Zhang, Z., Chen, N., Li, Z., et al. (2021). Assessing genomic diversity and signatures of selection in Jiaxian Red cattle using whole-genome sequencing data. *BMC Genomics* 22, 43–11. doi:10.1186/s12864-020-07340-0

Xie, S., Green, J., Beckers, J. F., and Roberts, R. M. (1995). The gene encoding bovine pregnancy-associated glycoprotein-1, an inactive member of the aspartic proteinase family. *Gene* 159, 193–197. doi:10.1016/0378-1119(94)00928-l

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:10.1016/j.ajhg.2010.11.011

Zhang, F., Wang, Y., Mukiibi, R., Chen, L., Vinsky, M., Plastow, G., et al. (2020). Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: Feed efficiency and component traits. *BMC Genomics* 21, 6362. doi:10.1186/S12864-019-6362-1

Zhang, Q., Guldbrandtsen, B., Bosse, M., Lund, M. S., and Sahana, G. (2015). Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics* 16, 542. doi:10.1186/s12864-015-1715-x

Zhang, S., Yao, Z., Li, X., Zhang, Z., Liu, X., Yang, P., et al. (2022). Assessing genomic diversity and signatures of selection in Pinan cattle using whole-genome sequencing data. *BMC Genomics* 23, 460–510. doi:10.1186/s12864-022-08645-y

Zhong, J. L., Xu, J. W., Wang, J., Wenfan, Y. H., Zheng, L., et al. (2019). A novel SNP of PLAG1 gene and its association with growth traits in Chinese cattle. *Gene* 689, 166–171. doi:10.1016/j.gene.2018.12.018

frontiers | Frontiers in Genetics

TYPE Perspective
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1010456

Evolutionary genomics: Insights from the invasive European starlings

Katarina C. Stuart[1]*, William B. Sherwin[1], Richard J. Edwards[2†] and Lee A Rollins[1†]

[1]Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, UNSW Sydney, Sydney, NSW, Australia, [2]Evolution & Ecology Research Centre, School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney, NSW, Australia

Check for updates

OPEN ACCESS

EDITED BY
Samuel A. Cushman,
Forest Service (USDA), United States

REVIEWED BY
Yuanda Lv,
Jiangsu Academy of Agricultural Sciences (JAAS), China

*CORRESPONDENCE
Katarina C. Stuart,
Katarina.Stuart@unsw.edu.au

†These authors share senior authorship

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

RECEIVED 03 August 2022
ACCEPTED 23 November 2022
PUBLISHED 04 January 2023

CITATION
Stuart KC, Sherwin WB, Edwards RJ and Rollins LA (2023), Evolutionary genomics: Insights from the invasive European starlings.
Front. Genet. 13:1010456.
doi: 10.3389/fgene.2022.1010456

Two fundamental questions for evolutionary studies are the speed at which evolution occurs, and the way that this evolution may present itself within an organism's genome. Evolutionary studies on invasive populations are poised to tackle some of these pressing questions, including understanding the mechanisms behind rapid adaptation, and how it facilitates population persistence within a novel environment. Investigation of these questions are assisted through recent developments in experimental, sequencing, and analytical protocols; in particular, the growing accessibility of next generation sequencing has enabled a broader range of taxa to be characterised. In this perspective, we discuss recent genetic findings within the invasive European starlings in Australia, and outline some critical next steps within this research system. Further, we use discoveries within this study system to guide discussion of pressing future research directions more generally within the fields of population and evolutionary genetics, including the use of historic specimens, phenotypic data, non-SNP genetic variants (e.g., structural variants), and pan-genomes. In particular, we emphasise the need for exploratory genomics studies across a range of invasive taxa so we can begin understanding broad mechanisms that underpin rapid adaptation in these systems. Understanding how genetic diversity arises and is maintained in a population, and how this contributes to adaptability, requires a deep understanding of how evolution functions at the molecular level, and is of fundamental importance for the future studies and preservation of biodiversity across the globe.

KEYWORDS

rapid adaptation, population genetics, *Sturnus vulgaris*, structural variants, plasticity, museum specimens

Introduction

Evolutionary theory states that the immense diversity existing on this planet does so through a complex combination of factors. These factors include genetics, epigenetics, and plasticity, and it is the interplay of these processes that allow species to evolve within our increasingly changing world (Goudie, 2018). We are ever gaining an appreciation for both the speed at which evolution occurs (Prentis et al., 2008), and the role humanity plays in shaping it (Sih et al., 2011). Thus there is much interest in the central role genetic variation plays in facilitating a population's evolutionary potential, so that we may better understand why some persist and others perish.

Population and evolutionary genomics play pivotal roles in answering these questions. Technical advances have given rise to cheap "reduced representation" data (subsampling genomic variation) and the growing feasibility of whole genome resequencing (WGS) for non-model organisms. Population genetics can now move beyond the characterisation of broad genetic patterns to uncover evolutionary important genetic variants at a resolution previously inaccessible to non-model organism studies (Hendricks et al., 2018; Hohenlohe et al., 2019, 2021). Examining genetic patterns across, for example, environmental (e.g., Gugger et al., 2017) or morphological (e.g., Nannan et al., 2022) landscapes, enables us to develop hypothesis regarding the drivers of a population or species' genetics.

Studies in invasive species genomics are fundamental to these efforts. By examining how invasive populations' genetic diversity is shaped by novel selection regimes, we obtain insight into the molecular patterns underpinning evolution. Invasive species, by nature, are successful following genetic bottlenecks (i.e., the large reduction in effective population size that occurs during translocation), and provide an avenue for understanding what aspects of genetic diversity (e.g., transposable elements; Stapley et al., 2015, or specific chromosomes; Meisel & Connallon 2013; Waters et al., 2021) contribute to adaptation under a new selection regime. Through such studies, we begin to appreciate the complex nature of genetic variation and how this may facilitate rapid local adaptation, and thereby species persistence, in response to an altered environment under a future of climate change (Razgour et al., 2019; Waldvogel et al., 2020). In this perspective, we discuss recent discoveries in genetics of the invasive European starlings within Australia, and use these studies to prompt interesting avenues for further research more broadly across the fields of evolutionary and population genetics.

Perspectives from the study of the European starling

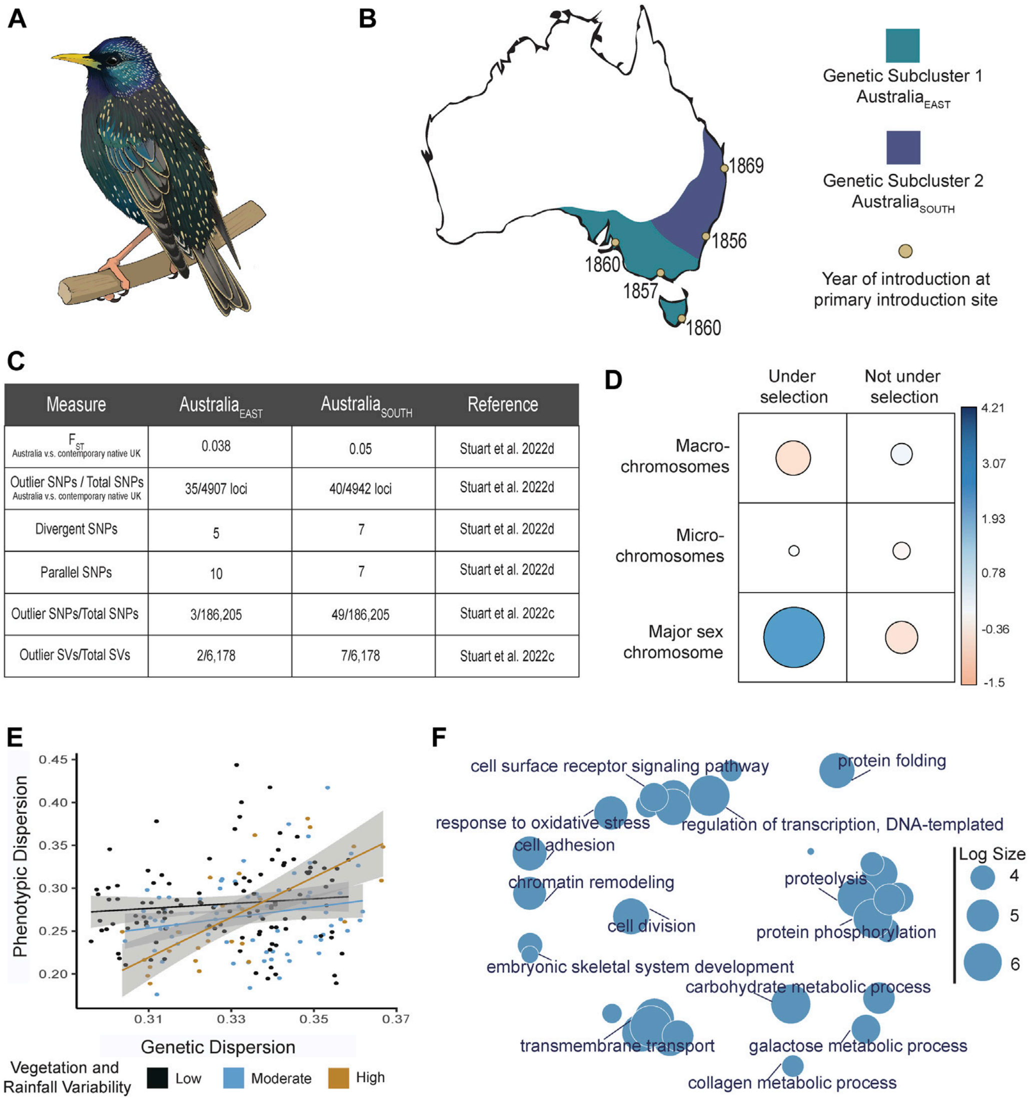
Of the 17,000 species that have been labelled as invasive (Seebens et al., 2017), the European starling (*Sturnus vulgaris*) (Figure 1A) is a standout. As one of the only birds on the IUCN's top 100 worst invasive species (Lowe et al., 2000), the starling, despite suffering dramatic population declines within its native palearctic range (Bowler et al., 2019), has colonised every other continent, barring Antarctica. The repeated and well documented introductions, combined with extensive natural history, genetic, and other biologically relevant data (e.g., environmental, phenotypic), has and will continue to yield many exciting discoveries in molecular evolution. With recent publication of several genetic studies on the invasive starling populations (focused primarily on Australia) comes an opportunity to synthesise key results, and propose broad hypothesis regarding the nature of their rapid evolution in native and introduced populations.

Complex introduction histories in human-mediated populations impact selection analysis

The Australian starling invasion has a well-documented, but complex, introduction history, with multiple geographically dispersed introduction points that are separated in some instances by thousands of kilometres (Figure 1B). Starling population genetics has demonstrated that while invasive populations provide a valuable resource for evolutionary studies, we must also acknowledge the challenges of separating selective and neutral evolutionary processes for populations with complex introduction histories. Demographic processes during range expansion may create false signals of selection in genetic data, or may even mask legitimate signals (Stuart et al., 2021). Further, a continuous invasive range may have resulted from numerous separate introductions that themselves were exposed to different selection regimes (Figure 1C), either at the introduction site or along environmental clines during range expansion (e.g., Stuart et al., 2022a). This selection co-occurs alongside source population differences and stochastic demographic effects (e.g., drift), hence it may be impossible to separate some genuine signals of selection from neutral genetic processes within some invasions. Strong subpopulation structure and recent range expansion may confound local signatures of adaptation, and therefore analytical approaches should take this into account (e.g., Stuart et al., 2022d; 2022c). The impact of separate introduction sites or range expansions on population-wide genetic variation is something that may need to be considered not just within invasive populations, but during genomic studies on non-invasive populations (e.g., Drury et al., 2017; Drinan et al., 2018; Pertoldi et al., 2021), and species reintroductions (e.g., Kaulfuss & Reisch 2017; Dincă et al., 2018; Mims et al., 2019).

Patterns of evolutionary change within the *Sturnus vulgaris* genome

Exploration of starling populations identified several broad trends that characterize patterns of genetic change in this species. First, comparing invasive populations on separate continents (Hofmeister et al., 2021), or distinct subpopulations within a



Measure	Australia$_{EAST}$	Australia$_{SOUTH}$	Reference
F_{ST} Australia v.s. contemporary native UK	0.038	0.05	Stuart et al. 2022d
Outlier SNPs / Total SNPs Australia v.s. contemporary native UK	35/4907 loci	40/4942 loci	Stuart et al. 2022d
Divergent SNPs	5	7	Stuart et al. 2022d
Parallel SNPs	10	7	Stuart et al. 2022d
Outlier SNPs/Total SNPs	3/186,205	49/186,205	Stuart et al. 2022c
Outlier SVs/Total SVs	2/6,178	7/6,178	Stuart et al. 2022c

FIGURE 1
Summary diagram of evolutionary trends for *Sturnus vulgaris* (*S. vulgaris*) within the invasive Australian range. Panel **(A)** depicts an artist's image of a male *S. vulgaris* in breeding season. Panel **(B)** depicts the Australian range of *S. vulgaris* (approximately based on eBird data retrieved 2018), with approximate genetic sub-structuring indicated in purple (Australia$_{EAST}$) and blue (Australia$_{SOUTH}$), and with introduction sites indicated (yellow circles) next to the year of first introduction. Panel **(C)** depicts differences in genetic differentiation between the two Australian subpopulations (Australia$_{EAST}$ and Australia$_{SOUTH}$) and the native range. F_{ST} values were obtained from comparisons to Newcastle, UK, from Stuart et al. (2022d). Panel **(D)** depicts a subset of the chi-squared results that assessed the occurrence of putative outliers and non-outlier SNPs across macro-, micro-, and major sex (Z) chromosome from Stuart et al. (2022d). The table visualises the Pearson residuals, where the circle area is proportional to the amount of the cell contribution, positive residuals (indicating a positive correlation) are in blue, and negative residuals (indicating a negative correlation) are in orange. Panel **(E)** depicts the positive interaction between phenotypic and genetic dispersion, which is positively affected by the level of ground cover vegetation and annual precipitation variation, data pulled from Stuart et al. (2022a). Panel **(F)** depicts a REVIGO (Supek et al., 2011) gene ontology term summary plot of all coding regions that were identified as significant across loci identified in studies on the Australian starling population: all statistically-outlier and environmentally-associated loci (Stuart et al., 2021); all divergent and parallel loci (Stuart et al., 2022d); all statistically outlier SNPs and SVs identified by BAYESCAN v2.1 (Foll and Gaggiotti, 2008) (Stuart et al., 2022c); all morphologically-associated loci that were also under selection (Stuart et al., 2022a). Log size is indicative of the frequency of the GO terms.

single invasive range (Stuart et al., 2022d), reveals that separate populations may experience parallel selection across geographically isolated regions. Further, selection (parallel or divergent) is not restricted to translocated invasive range, but may also occur within the native range post divergence (Stuart et al., 2022d). Considering this, it is vital that evolutionary studies comparing putatively adaptive differences between populations do not assume either the origin of parallel signatures of selection, or that genetic divergence has occurred only or even primarily within invasive ranges, because this may limit interpretation of results.

Second, selection (parallel or divergent) within starling populations often occurs at sites with moderate allele frequencies (Stuart et al., 2022d), and levels of balancing selection (selection that maintains multiple allele variants within a population) vary across different types of genetic variants (Stuart et al., 2022c). This supports the theory that balancing selection within native populations has the potential to maintain evolutionarily important alleles (Hedrick, 2007) that may undergo directional selection within novel ranges, assisting an organisms' ability to rapidly adapt to new selection regimes (Stern and Lee, 2020). Understanding the mechanisms maintaining standing genetic variation within populations, particularly for functional variants of different types (e.g., single nucleotide polymorphisms—SNPs, structural variants—SVs) has important implications for conservation genetics (e.g., Whiteley et al., 2015; Lai et al., 2019) and thus is an extremely pressing research direction.

Third, examining the location of putative sites under selection across the starling's genome reveals a larger number of genomic sites than expected under selection in the major sex (Z) chromosome (Stuart et al., 2022d, Figure 1D). These results concur with existing research on the role that sex chromosomes play in rapid evolution (Meisel & Connallon 2013), and is in alignment with results that highlight the importance of the Z chromosome specifically within avian divergence and speciation (Campagna et al., 2017).

Lastly, assessing genetic alongside environmental and morphological data indicates that genetic variation is positively correlated with phenotypic variation, and this relationship becomes more pronounced under a temporally variable environment (high vegetation and rainfall variability; Stuart et al., 2022a, Figure 1E). While the exact details of these results (e.g., driving climate factors) are system specific, they demonstrate the importance of assessing variation across data types to better understand drivers of phenotypic plasticity (which may in the future be aided by the inclusion of epigenome data in plasticity studies). Importantly, because climate change increases environmental variability (Thornton et al., 2014), such studies on invasive populations will allow us to understand how genetics and environments, and their interactions, shape rapid adaptive change under novel selection regimes.

Signals of selection correlated with environmental variables

Strong genetic differences between the invasive Australian and North American starling populations (Hofmeister et al., 2021; Stuart et al., 2021) showcase the flexible ecology of this globally invasive species, demonstrating the difficulties in predicting potential species' ranges (Whitney and Gabler 2008). A focal research direction within recent starling genomic studies sought to understand how the hotter, more arid Australian environment shaped the introduced starling population. Within the two major Australian subpopulations, the southern subpopulation cluster is more divergent from the native range than the eastern one (Stuart et al., 2022c, 2022d, key results summarised in Figure 1C). While these patterns may have resulted from founder genetics or stochastic processes, they may also be indicative of different selection regimes experienced within subpopulations. There are strong bottleneck effects at the Australian western-most range-edge, along with high proportions of private alleles for both SNPs and SVs (Rollins et al., 2011; Stuart et al., 2021; 2022c). Likely, this range-edge differentiation (previously attributed to an extreme genetic bottleneck) resulted from accidental introduction/s from elsewhere in the starlings' native range.

These studies have flagged many coding regions under putative selection within the invasive Australian range (Figure 1F). These contain genes covering a range of putatively adaptive functions (e.g., immune response, beak morphology), including a diverse range of novel seeking behaviour associated genes (Stuart et al., 2021; Stuart et al., 2022c, Stuart et al., 2022d, but see Rollins et al., 2015). Starlings' intelligence and behavior has been the focus of much research (Mueller et al., 2014; Nettle et al., 2015; Van Berkel et al., 2018), and with research linking invasion success to behavioral flexibility (Sol et al., 2002), follow-up studies on these candidate genes will shed light on rapid adaptation in novel seeking behavior within starlings.

Rapid adaptation, occasionally in response to environment, has impacted starling phenotype. Analysis of morphology-associated loci under putative selection indicated that climate extremes are more important than means in explaining morphological patterns (Stuart et al., 2022a), echoing findings within other species (e.g., Vasseur et al., 2014; Gardner et al., 2017). Further, climate correlates of genetic patterns were highly varied across overall genetic patterns, morphology-associated genetic patterns, and phenotype and genetic variance patterns. Collectively, these results demonstrate how important it is to examine different components of a genetic landscape to best understand what is driving adaptive change.

Areas for future expansion in population and evolutionary genomics

Synthesis of recent findings in starling genetics has identified important growing themes and promising future directions in population and evolutionary genomics research. These avenues will provide invaluable insights into the fundamentals of genomic evolution, with application in both invasive and non-invasive systems.

The utility of museum collections

Sequencing historical samples from museum collections facilitates a range of potential new genomic projects to, for example, track temporal changes in allelic landscapes, or conduct studies into now-extinct lineages. Analysis alongside contemporary samples in both our study and others is promising (e.g., Ewart et al., 2019; Parejo et al., 2020; Yao et al., 2020). Across many institutions, research is ongoing into different aspects of museum sample utility, including extraction methods (Tsai et al., 2020; Hahn et al., 2022), DNA recovery from specimen ethanol (Jeunen et al., 2021), and improving wet lab—bioinformatic hybrid approaches (Bernstein and Ruane, 2022) to maximise data from rare and degraded specimens (Raxworthy and Smith, 2021). Assessing, for example, morphological and genetic change over time, may be used to better appreciate how industrialization, human land use, and climate change has affected a wide range of species. Conducting such studies across both successful invasives and vulnerable geographically-isolated endemics would facilitate deeper understanding of how population expansion (or decline) tracks with underlying genetic diversity and anthropogenic effects.

The need for collection of phenotypic data alongside genetics

While phenotypic data is often well integrated into genetic studies in agricultural species (e.g., Reynolds et al., 2021), managed native species (e.g., Álvarez-Varas et al., 2021), or even plant invasions (e.g., Bhattarai et al., 2017), there is a general lack of such data in invasive animal studies. Even when phenotype data are collected, sampling wild populations means that some data remains unknown (e.g., pedigree information) and many traditional analytical techniques (e.g., heritability analysis) may require unobtainably large sample sizes. And while museum collections may enable easy morphological data collection, such collection efforts must contend with preservation method related shrinkage (Maayan et al., 2022). However, these difficulties should be and are being overcome (e.g., Hedrick et al., 2018), because pairing phenotype data with the underlying genetic data is vital for understanding the role plasticity plays in invasions (e.g., Santi et al., 2020), as well as identifying critical genome regions that may facilitate rapid phenotypic adaptation (e.g., Wu et al., 2019). Understanding heritability and plasticity is necessary for long-term modelling of invasive populations, and to explore the limits of a species' ability to adapt to shifting selection regimes.

Beyond the genome: A multi-omics approach to adaptation

General consideration of phenotypic plasticity is important for establishing the limits of genetic (evolved) contributions to adaptation. However, phenotypic plasticity itself can have different underlying causes and mechanisms. Deeper understanding the role of plasticity in shaping a populations' adaptive potential requires expanding omics data collection beyond just the genome. Co-analysing genetic and phenotypic data will be greatly aided by the inclusion of, for example, transcriptome, proteome, or epigenome data (Layton and Bradbury, 2022). These complimentary data sets will provide vital information about the biological processes that link the underlying heritable DNA of an organism to its resulting phenotype. Exciting new techniques are allowing us to obtain this information from previously inaccessible samples (e.g., Hahn et al., 2020; Rubi et al., 2020). Analysing multi-omics data sets across temporal and spatial landscapes will shed light onto the complex interactions between the heritable genetic, heritable epigenetic, and non-heritable plastic elements that collectively contribute to a populations' adaptive potential.

Putatively adaptive loci

Evolutionary genomic studies often produce a list of genetic sites flagged as under putative selection or associated with phenotypic or environmental data. While generating a shortlist of biologically interesting variants is the first of many steps towards biologically validating these results, knowing whether a variant has an adaptive advantage (or disadvantage) may require additional data (e.g., RNA expression or phenotype data), or even "evolve and resequence" studies under one or more standard environmental conditions (Schlötterer et al., 2015). It is vital that we go beyond compilation of outlier loci and begin using this often end-result as a stepping stone to further scientific inquiry. Analytical advancements, such as Alphafold (Jumper et al., 2021) and Variant Effect Predictor (McLaren et al., 2016), will help evolutionary biologists ask more of their data, and reciprocally increase our understanding of science across research fields. We may begin to use flagged outlier loci across a range of species to examine broad questions about, for example, genomic positioning of selective loci, protein sequence change trends, and 3D protein structure impacts. Additionally, for well-studied systems like the starling, it may be worthwhile to curate species-specific databases of variants of interest, as often genes of import are accessible only by manually trawling through literature, and non-coding flagged loci

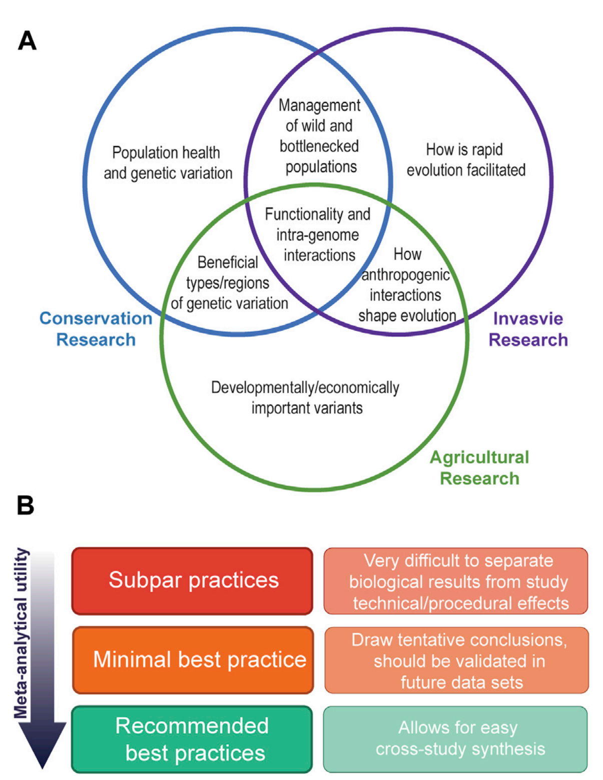

A
Population health and genetic variation
Management of wild and bottlenecked populations
How is rapid evolution facilitated
Functionality and intra-genome interactions
Beneficial types/regions of genetic variation
How anthropogenic interactions shape evolution
Conservation Research
Invasvie Research
Developmentally/economically important variants
Agricultural Research
B
Meta-analytical utility
Subpar practices
Very difficult to separate biological results from study technical/procedural effects
Minimal best practice
Draw tentative conclusions, should be validated in future data sets
Recommended best practices
Allows for easy cross-study synthesis

FIGURE 2
Visualisation of the possible outcomes and data requirements of cross study synthesis in evolutionary and population genetics. Panel **(A)** depicts common research outcomes from studies of invasive, conservation, and agricultural population and evolutionary genomics research, and their overlap. Panel **(B)** depicts the need for highly curated and reproducible datasets to enable confident cross-study comparisons. Best practices will often be specific to the variant types being called and the research setting they are being conducted in (e.g., Olson et al., 2015; Cameron et al., 2019; Koboldt 2020), and will often be limited by funding and resource availability. In general, minimal best practices should be informed by both the technical limitations of the sequence technology used (e.g., quality checks of the right type and stringency), as well as biologically informed decision making (e.g., accounting for population structuring when conducting downstream analysis such as outlier detection).

are even more inaccessible (if recorded at all). Compiling outlier variant data across studies is vital if we want to develop inter-specific perspectives. Ultimately, through clear scientific reporting of the results of individual exploratory studies we may begin to collectively resolve broad trends in molecular evolution.

Genomic meta-analysis and whole genome resequencing

As we acknowledge the utility in cross-study comparisons of variants of interest, so too must we acknowledge the lack of these within the field of evolutionary and population genomics. Such approaches can start more modestly, with comparisons across similar species (for example, comparison between the invasive sturnids, the starling and the common myna *Acridotheres tristis*) to yield insights into the repeatability of adaptive change in invasive taxa. While the lack of comparative studies is simply because the field is in its relative infancy, we are quickly moving into an age where enough studies have now been published for us to begin employing meta-analytical approaches to conduct formal inter -population or -specific comparisons. This will allow us to begin confirming whether results from a singular study are an isolated phenomenon within a particular system, or are broadly applicable over many. Comparing patterns of genomic change across a broad range of invasive species is of vital importance to invasion population genomics to answer questions about the importance of different genomic variants and the role they play in rapid adaptation and species persistence. Further, comparisons between invasion genomics and the fields of conservation and agricultural genomics also promise to yield interesting answers (Figure 2A).

Technology continues to shape the nature of the questions that can be asked and reveals the importance of previously understudied types of variation. WGS captures genome-wide genetic information that improves on reduced representation sequencing approaches, as it does not rely on subsampling species-specific genomic loci. This provides a more uniform starting point for further analysis and more long-term utility, provided careful scientific reporting and best practices are followed (Figure 2B). Further, WGS has broadened investigation of non-SNP genetic variants, such as SVs and transposable elements, which are of growing interest to the evolution and population genomics community. These investigations are increasingly served by shifts to “third generation” technologies and the promise of affordable, high accuracy long-read sequencing.

Pan-genomes

The incorporation of published data into cross study comparisons is greatly aided by a high-quality species genome. While putative chromosomes may assembled using syntenic approaches (e.g., Stuart et al., 2022b), completely de-novo assemblies with long-range scaffolding data incorporated are necessary for confidence in, for example, structural rearrangements. However, even these platinum standard genomes are superseded by pan-genomes, which is a genome map that attempts to capture species-wide genetic diversity and structure (Vernikos et al., 2015; Sherman and Salzberg, 2020). Pan-genomes will help to alleviate reference biases introduced by mapping many, possibly quite genetical distinct individuals of a species to a reference genome that only represents the genetic structure of one individual (and thus may fail to provide an adequate map for more divergent regions within the resequenced individuals), increasing data integrity and hence utility (Figure 2B). While the creation of such a resource is currently not financially feasible for most studies, existing genomes can be incorporated into future pan-genomes. To fully characterize genetic divergence between populations, particularly in hard to map genomic areas (such as regions with high repeat content), pan-genomes are essential.

Conclusion

The study of invasive systems have and will continue to yield many important discoveries within the fields of population and evolutionary genomics. Characterising rapid adaptation from the molecular to the macro level within invasive populations enables scientists to appreciate how genetic variation interacts with a variety of selection processes to allow species to evolve. Within these studies, starlings present a valuable system for understanding evolution, providing opportunity to investigate everything from subtle morphological shifts within a region, to observing broad patterns of parallel change across the globe. The research conducted on this system demonstrates the complex nature of standing genetic diversity and the means through which it facilitates adaptation. From these results we also see some promising directions of future study within the field of evolutionary and population genomics. Collectively pursuing these directions using this system and that of other invaders will facilitate a deeper appreciation of how evolution functions at the molecular level. Ultimately, through deep studies across a broad range of taxa, we may learn to precisely explain and predict patterns of evolution, to protect precious biodiversity.

Ethics statement

Ethical review and approval was not required for the animal study because samples and data were pre-existing and used in other projects.

Author contributions

KS, RE, and LR conceived this project. Writing and editing was completed by all authors. Figures were created by KS.

Funding

LR was supported by a Scientia Fellowship from UNSW.

Acknowledgments

Megan Bishop created the starling illustration in Figure 1. We thank the reviewer for comments that improved this manuscript.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Álvarez-Varas, R., Heidemeyer, M., Riginos, C., Benítez, H. A., Reséndiz, E., Lara-Uc, M., et al. (2021). Integrating morphological and genetic data at different spatial scales in a cosmopolitan marine turtle species: Challenges for management and conservation. *Zool. J. Linn. Soc.* 191, 434–453. doi:10.1093/zoolinnean/zlaa066

Bernstein, J. M., and Ruane, S. (2022). Maximizing molecular data from low-quality fluid-preserved specimens in natural history collections. *Front. Ecol. Evol.* 10, 893088. doi:10.3389/fevo.2022.893088

Bhattarai, G. P., Meyerson, L. A., Anderson, J., Cummings, D., Allen, W. J., and Cronin, J. T. (2017). Biogeography of a plant invasion: Genetic variation and plasticity in latitudinal clines for traits related to herbivory. *Ecol. Monogr.* 87, 57–75. doi:10.1002/ecm.1233

Bowler, D. E., Heldbjerg, H., Fox, A. D., de Jong, M., and Böhning-Gaese, K. (2019). Long-term declines of European insectivorous bird populations and potential causes. *Conserv. Biol.* 33, 1120–1130. doi:10.1111/cobi.13307

Cameron, D. L., Di Stefano, L., and Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.* 10, 3240. doi:10.1038/s41467-019-11146-4

Campagna, L., Repenning, M., Silveira, L. F., Fontana, C. S., Tubaro, P. L., and Lovette, I. J. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. *Sci. Adv.* 3, e1602404. doi:10.1126/sciadv.1602404

Dincă, V., Bálint, Z., Vodă, R., Dapporto, L., Hebert, P. D. N., and Vila, R. (2018). Use of genetic, climatic, and microbiological data to inform reintroduction of a regionally extinct butterfly. *Conserv. Biol.* 32, 828–837. doi:10.1111/cobi.13111

Drinan, D. P., Gruenthal, K. M., Canino, M. F., Lowry, D., Fisher, M. C., and Hauser, L. (2018). Population assignment and local adaptation along an isolation-by-distance gradient in Pacific cod (*Gadus macrocephalus*). *Evol. Appl.* 11, 1448–1464. doi:10.1111/eva.12639

Drury, C., Schopmeyer, S., Goergen, E., Bartels, E., Nedimyer, K., Johnson, M., et al. (2017). Genomic patterns in Acropora cervicornis show extensive population structure and variable genetic diversity. *Ecol. Evol.* 7, 6188–6200. doi:10.1002/ece3.3184

Ewart, K. M., Johnson, R. N., Ogden, R., Joseph, L., Frankham, G. J., and Lo, N. (2019). Museum specimens provide reliable SNP data for population genomic analysis of a widely distributed but threatened cockatoo species. *Mol. Ecol. Resour.* 19, 1578–1592. doi:10.1111/1755-0998.13082

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

Gardner, J. L., Rowley, E., de Rebeira, P., de Rebeira, A., and Brouwer, L. (2017). Effects of extreme weather on two sympatric Australian passerine bird species. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20160148. doi:10.1098/rstb.2016.0148

Goudie, A. S. (2018). *Human impact on the natural environment*. New York: John Wiley & Sons.

Gugger, P. F., Liang, C. T., Sork, V. L., Hodgskiss, P., and Wright, J. W. (2017). Applying landscape genomic tools to forest management and restoration of Hawaiian koa (Acacia koa) in a changing environment. *Evol. Appl.* 11, 231–242. doi:10.1111/eva.12534

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., and Holleley, C. E. (2022). Unlocking inaccessible historical genomes preserved in formalin. *Mol. Ecol. Resour.* 22, 2130–2147. doi:10.1111/1755-0998.13505

Hahn, E. E., Grealy, A., Alexander, M., and Holleley, C. E. (2020). Museum epigenomics: Charting the future by unlocking the past. *Trends Ecol. Evol.* 35, 295–300. doi:10.1016/j.tree.2019.12.005

Hedrick, B. P., Yohe, L., Vander Linden, A., Dávalos, L. M., Sears, K., Sadier, A., et al. (2018). Assessing soft-tissue shrinkage estimates in museum specimens imaged with diffusible iodine-based contrast-enhanced computed tomography (diceCT). *Microsc. Microanal.* 24, 284–291. doi:10.1017/S1431927618000399

Hedrick, P. W. (2007). Balancing selection. *Curr. Biol.* 17, R230–R231. doi:10.1016/j.cub.2007.01.012

Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., et al. (2018). Recent advances in conservation and population genomics data analysis. *Evol. Appl.* 11, 1197–1211. doi:10.1111/eva.12659

Hofmeister, N. R., Stuart, K., Warren, W. C., Werner, S. J., Bateson, M., Ball, G. F., et al. (2021). Concurrent invasions by European starlings (*Sturnus vulgaris*) suggest selection on shared genomic regions even after genetic bottlenecks. 2021.05.19.442026. doi:10.1101/2021.05.19.442026

Hohenlohe, P. A., Funk, W. C., and Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Mol. Ecol.* 30, 62–82. doi:10.1111/mec.15720

Hohenlohe, P. A., Hand, B. K., Andrews, K. R., and Luikart, G. (2019). "Population genomics provides key insights in ecology and evolution," in *Population genomics: Concepts, approaches and applications population genomics*. Editor O. P. Rajora (Cham: Springer International Publishing), 483–510. doi:10.1007/13836_2018_20

Jeunen, G.-J., Cane, J. S., Ferreira, S., Strano, F., Ammon, U., Cross, H., et al. (2021). Assessing the utility of marine filter feeders for environmental DNA (eDNA) biodiversity monitoring. 2021.12.21.473722. doi:10.1101/2021.12.21.473722

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kaulfuss, F., and Reisch, C. (2017). Reintroduction of the endangered and endemic plant species cochlearia bavarica—implications from conservation genetics. *Ecol. Evol.* 7, 11100–11112. doi:10.1002/ece3.3596

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Med.* 12, 91. doi:10.1186/s13073-020-00791-w

Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., et al. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2152–2157. doi:10.1073/pnas.1813597116

Layton, K. K. S., and Bradbury, I. R. (2022). Harnessing the power of multi-omics data for predicting climate change response. *J. Anim. Ecol.* 91, 1064–1072. doi:10.1111/1365-2656.13619

Lowe, S., Browne, M., and Boudjelas, S. (2000). *100 of the world's worst invasive alien species. A selection from the global invasive species database*. Auckland, New Zealand: Invasive Species Specialist Group.

Maayan, I., Reynolds, R. G., Goodman, R. M., Hime, P. M., Bickel, R., Luck, E. A., et al. (2022). Fixation and preservation contribute to distortion in vertebrate museum specimens: A 10-year study with the lizard Anolis sagrei. *Biol. J. Linn. Soc.* 136, 443–454. doi:10.1093/biolinnean/blac040

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17, 122. doi:10.1186/s13059-016-0974-4

Meisel, R. P., and Connallon, T. (2013). The faster-X effect: Integrating theory and data. *Trends Genet.* 29, 537–544. doi:10.1016/j.tig.2013.05.009

Mims, M. C., Day, C. C., Burkhart, J. J., Fuller, M. R., Hinkle, J., Bearlin, A., et al. (2019). Simulating demography, genetics, and spatially explicit processes to inform reintroduction of a threatened char. *Ecosphere* 10, e02589. doi:10.1002/ecs2.2589

Mueller, J. C., Edelaar, P., Carrete, M., Serrano, D., Potti, J., Blas, J., et al. (2014). Behaviour-related DRD4 polymorphisms in invasive bird populations. *Mol. Ecol.* 23, 2876–2885. doi:10.1111/mec.12763

Nannan, L., Huamiao, L., Yan, J., Xingan, L., Yang, L., Tianjiao, W., et al. (2022). Geometric morphology and population genomics provide insights into the adaptive evolution of *Apis cerana* in Changbai Mountain. *BMC Genomics* 23, 64. doi:10.1186/s12864-022-08298-x

Nettle, D., Andrews, C. P., Monaghan, P., Brilot, B. O., Bedford, T., Gillespie, R., et al. (2015). Developmental and familial predictors of adult cognitive traits in the European starling. *Anim. Behav.* 107, 239–248. doi:10.1016/j.anbehav.2015.07.002

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., et al. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet.* 6, 235. doi:10.3389/fgene.2015.00235

Parejo, M., Wragg, D., Henriques, D., Charrière, J.-D., and Estonba, A. (2020). Digging into the genomic past of Swiss honey bees by whole-genome sequencing museum specimens. *Genome Biol. Evol.* 12, 2535–2551. doi:10.1093/gbe/evaa188

Pertoldi, C., Ruiz-Gonzalez, A., Bahrndorff, S., Renee Lauridsen, N., Nisbeth Henriksen, T., Eskildsen, A., et al. (2021). Strong isolation by distance among local populations of an endangered butterfly species (Euphydryas aurinia). *Ecol. Evol.* 11, 12790–12800. doi:10.1002/ece3.8027

Prentis, P. J., Wilson, J. R. U., Dormontt, E. E., Richardson, D. M., and Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends Plant Sci.* 13, 288–294. doi:10.1016/j.tplants.2008.03.004

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi:10.1016/j.tree.2021.07.009

Razgour, O., Forester, B., Taggart, J. B., Bekaert, M., Juste, J., Ibáñez, C., et al. (2019). Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *Proc. Natl. Acad. Sci. U. S. A.* 116, 10418–10423. doi:10.1073/pnas.1820663116

Reynolds, E. G., Neeley, C., Lopdell, T. J., Keehan, M., Dittmer, K., Harland, C. S., et al. (2021). Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. *Nat. Genet.* 53, 949–954. doi:10.1038/s41588-021-00872-5

Rollins, L. A., Whitehead, M. R., Woolnough, A. P., Sinclair, R., and Sherwin, W. B. (2015). Is there evidence of selection in the dopamine receptor D4 gene in Australian invasive starling populations? *Curr. Zool.* 61, 505–519. doi:10.1093/czoolo/61.3.505

Rollins, L. A., Woolnough, A. P., Sinclair, R., Mooney, N. J., and Sherwin, W. B. (2011). Mitochondrial DNA offers unique insights into invasion history of the common starling. *Mol. Ecol.* 20, 2307–2317. doi:10.1111/j.1365-294X.2011.05101.x

Rubi, T. L., Knowles, L. L., and Dantzer, B. (2020). Museum epigenomics: Characterizing cytosine methylation in historic museum specimens. *Mol. Ecol. Resour.* 20, 1161–1170. doi:10.1111/1755-0998.13115

Santi, F., Riesch, R., Baier, J., Grote, M., Hornung, S., Jüngling, H., et al. (2020). A century later: Adaptive plasticity and rapid evolution contribute to geographic variation in invasive mosquitofish. *Sci. Total Environ.* 726, 137908. doi:10.1016/j.scitotenv.2020.137908

Schlötterer, C., Kofler, R., Versace, E., Tobler, R., and Franssen, S. U. (2015). Combining experimental evolution with next-generation sequencing: A powerful tool to study adaptation from standing genetic variation. *Heredity* 114, 431–440. doi:10.1038/hdy.2014.86

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., et al. (2017). No saturation in the accumulation of alien species worldwide. *Nat. Commun.* 8, 14435. doi:10.1038/ncomms14435

Sherman, R. M., and Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nat. Rev. Genet.* 21, 243–254. doi:10.1038/s41576-020-0210-7

Sih, A., Ferrari, M. C. O., and Harris, D. J. (2011). Evolution and behavioural responses to human-induced rapid environmental change. *Evol. Appl.* 4, 367–387. doi:10.1111/j.1752-4571.2010.00166.x

Sol, D., Timmermans, S., and Lefebvre, L. (2002). Behavioural flexibility and invasion success in birds. *Anim. Behav.* 63, 495–502. doi:10.1006/anbe.2001.1953

Stapley, J., Santure, A. W., and Dennis, S. R. (2015). Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol. Ecol.* 24, 2241–2252. doi:10.1111/mec.13089

Stern, D. B., and Lee, C. E. (2020). Evolutionary origins of genomic adaptations in an invasive copepod. *Nat. Ecol. Evol.* 4, 1084–1094. doi:10.1038/s41559-020-1201-y

Stuart, K. C., Cardilini, A. P. A., Cassey, P., Richardson, M. F., Sherwin, W. B., Rollins, L. A., et al. (2021). Signatures of selection in a recent invasion reveal adaptive divergence in a highly vagile invasive species. *Mol. Ecol.* 30, 1419–1434. doi:10.1111/mec.15601

Stuart, K. C., Cardilini, A. P. A., Sherwin, W. B., and Rollins, L. A. (2022a). Genetics and plasticity are responsible for ecogeographical patterns in a recent invasion. *Front. Genet.* 13, 824424. doi:10.3389/fgene.2022.824424

Stuart, K. C., Edwards, R. J., Cheng, Y., Warren, W. C., Burt, D. W., Sherwin, W. B., et al. (2022b). Transcript- and annotation-guided genome assembly of the European starling. *Mol. Ecol. Resour.* 22, 3141–3160. doi:10.1111/1755-0998.13679

Stuart, K. C., Edwards, R. J., Sherwin, W. B., and Rollins, L. A. (2022c). Contrasting patterns of single nucleotide polymorphisms and structural variations across multiple invasions. 2022.07.04.498653. doi:10.1101/2022.07.04.498653

Stuart, K. C., Sherwin, W. B., Austin, J. J., Bateson, M., Eens, M., Brandley, M. C., et al. (2022d). Historical museum samples enable the examination of divergent and parallel evolution during invasion. *Mol. Ecol.* 31, 1836–1852. doi:10.1111/mec.16353

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* 6, e21800. doi:10.1371/journal.pone.0021800

Thornton, P. K., Ericksen, P. J., Herrero, M., and Challinor, A. J. (2014). Climate variability and vulnerability to climate change: A review. *Glob. Change Biol.* 20, 3313–3328. doi:10.1111/gcb.12581

Tsai, W. L. E., Schedl, M. E., Maley, J. M., and McCormack, J. E. (2020). More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. *Mol. Ecol. Resour.* 20, 1220–1227. doi:10.1111/1755-0998.13077

Van Berkel, M., Bateson, M., Nettle, D., and Dunn, J. (2018). Can starlings use a reliable cue of future food deprivation to adaptively modify foraging and fat reserves? *Anim. Behav.* 142, 147–155. doi:10.1016/j.anbehav.2018.06.015

Vasseur, D. A., DeLong, J. P., Gilbert, B., Greig, H. S., Harley, C. D. G., McCann, K. S., et al. (2014). Increased temperature variation poses a greater risk to species than climate warming. *Proc. R. Soc. B Biol. Sci.* 281, 20132612. doi:10.1098/rspb.2013.2612

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154. doi:10.1016/j.mib.2014.11.016

Waldvogel, A.-M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., et al. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evol. Lett.* 4, 4–18. doi:10.1002/evl3.154

Waters, P. D., Patel, H. R., Ruiz-Herrera, A., Álvarez-González, L., Lister, N. C., Simakov, O., et al. (2021). Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2112494118. doi:10.1073/pnas.2112494118

Whitney, K. D., and Gabler, C. A. (2008). Rapid evolution in introduced species, 'invasive traits' and recipient communities: challenges for predicting invasive potential. *Divers. Distrib.* 14, 569–580. doi:10.1111/j.1472-4642.2008.00473.x

Whiteley, A. R., Fitzpatrick, S. W., Funk, W. C., and Tallmon, D. A. (2015). Genetic rescue to the rescue. *Trends Ecol. Evol.* 30, 42–49. doi:10.1016/j.tree.2014.10.009

Wu, N., Zhang, S., Li, X., Cao, Y., Liu, X., Wang, Q., et al. (2019). Fall webworm genomes yield insights into rapid adaptation of invasive species. *Nat. Ecol. Evol.* 3, 105–115. doi:10.1038/s41559-018-0746-5

Yao, L., Witt, K., Li, H., Rice, J., Salinas, N. R., Martin, R. D., et al. (2020). Population genetics of wild *Macaca fascicularis* with low-coverage shotgun sequencing of museum specimens. *Am. J. Phys. Anthropol.* 173, 21–33. doi:10.1002/ajpa.24099

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 27 February 2023
DOI 10.3389/fgene.2023.1041103

Population structure and evolutionary history of the greater cane rat (*Thryonomys swinderianus*) from the Guinean Forests of West Africa

Isaac A. Babarinde[1,2†], Adeniyi C. Adeola[3,4,5*†], Chabi A. M. S. Djagoun[6†], Lotanna M. Nneji[7†], Agboola O. Okeyoyin[8], George Niba[9], Ndifor K. Wanzie[10,11], Ojo C. Oladipo[12], Ayotunde O. Adebambo[13], Semiu F. Bello[14], Said I. Ng'ang'a[3], Wasiu A. Olaniyi[15], Victor M. O. Okoro[16], Babatunde E. Adedeji[17], Omotoso Olatunde[17], Adeola O. Ayoola[3,4], Moise M. Matouke[18], Yun-yu Wang[19], Oscar J. Sanke[20], Saidu O. Oseni[21], Christopher D. Nwani[22] and Robert W. Murphy[23]

[1]Shenzhen Key Laboratory of Gene Regulation and Systems Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China, [2]Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China, [3]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, [4]Sino-Africa Joint Research Centre, Chinese Academy of Sciences, Kunming, China, [5]Centre for Biotechnology Research, Bayero University, Kano, Nigeria, [6]Laboratory of Applied Ecology, Faculty of Agronomic Sciences, University of Abomey-Calavi, Cotonou, Benin, [7]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, United States, [8]National Park Service Headquarters, Federal Capital Territory, Abuja, Nigeria, [9]National Centre for Animal Husbandry and Veterinary Training, Jakiri, North West Region, Cameroon, [10]Department of Zoology, University of Douala, Douala, Cameroon, [11]Department of Zoology, Faculty of Life Sciences, University of Ilorin, Ilorin, Kwara State, Nigeria, [12]Old Oyo National Park, Oyo, Nigeria, [13]Animal Genetics & Biotechnology, Federal University of Agriculture, Abeokuta, Nigeria, [14]Department of Animal Genetics, Breeding and Reproduction, College of Animal Science, South China Agricultural University, Guangzhou, China, [15]Department of Animal Science, Faculty of Agriculture, Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria, [16]Department of Animal Science and Technology, School of Agriculture and Agricultural Technology, Federal University of Technology, Owerri, Nigeria, [17]Department of Zoology, University of Ibadan, Ibadan, Oyo State, Nigeria, [18]Department of Fisheries and Aquatic Resources Management, University of Buea, Buea, Cameroon, [19]Wild Forensic Center, Kunming, China, [20]Taraba State Ministry of Agriculture and Natural Resources, Jalingo, Nigeria, [21]Department of Animal Sciences, Faculty of Agriculture, Obafemi Awolowo University, Ile-Ife, Nigeria, [22]Department of Zoology and Environmental Biology, Faculty of Biological Sciences, University of Nigeria, Nsukka, Nigeria, [23]Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, Toronto, ON, Canada

EDITED BY
Samuel A. Cushman,
Forest Service, United States Department of Agriculture (USDA), United States

REVIEWED BY
Giovanni Zecca,
Independent researcher, Milan, Italy
Luc Hippolyte Dossa,
University of Abomey-Calavi, Benin
Justice Opare Odoi,
Council for Scientific and Industrial Research (CSIR), Ghana

*CORRESPONDENCE
Adeniyi C. Adeola,
✉ chadeola@mail.kiz.ac.cn

†These authors have contributed equally to this work and share first authorship

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

RECEIVED 10 September 2022
ACCEPTED 07 February 2023
PUBLISHED 27 February 2023

CITATION
Babarinde IA, Adeola AC, Djagoun CAMS, Nneji LM, Okeyoyin AO, Niba G, Wanzie NK, Oladipo OC, Adebambo AO, Bello SF, Ng'ang'a SI, Olaniyi WA, Okoro VMO, Adedeji BE, Olatunde O, Ayoola AO, Matouke MM, Wang Y-y, Sanke OJ, Oseni SO, Nwani CD and Murphy RW (2023), Population structure and evolutionary history of the greater cane rat (*Thryonomys swinderianus*) from the Guinean Forests of West Africa. *Front. Genet.* 14:1041103.
doi: 10.3389/fgene.2023.1041103

Grasscutter (*Thryonomys swinderianus*) is a large-body old world rodent found in sub-Saharan Africa. The body size and the unique taste of the meat of this major crop pest have made it a target of intense hunting and a potential consideration as a micro-livestock. However, there is insufficient knowledge on the genetic diversity of its populations across African Guinean forests. Herein, we investigated the genetic diversity, population structures and evolutionary history of seven Nigerian wild grasscutter populations together with individuals from Cameroon, Republic of Benin, and Ghana, using five mitochondrial fragments, including D-loop and cytochrome b (*CYTB*). D-loop haplotype diversity ranged from 0.571 (± 0.149) in Republic of Benin to 0.921 (± 0.013) in Ghana. Within Nigeria, the haplotype diversity ranged from 0.659 (± 0.059) in

Cross River to 0.837 (± 0.075) in Ondo subpopulation. The fixation index (F_{ST}), haplotype frequency distribution and analysis of molecular variance revealed varying levels of population structures across populations. No significant signature of population contraction was detected in the grasscutter populations. Evolutionary analyses of *CYTB* suggests that South African population might have diverged from other populations about 6.1 (2.6–10.18, 95% CI) MYA. Taken together, this study reveals the population status and evolutionary history of grasscutter populations in the region.

KEYWORDS

genetic diversity, population structure, lower guinean forests, mitochondrial sequences, *Thryonomys swinderianus*

1 Introduction

Grasscutter or greater cane rat (*Thryonomys swinderianus*) is one of the two known extant cane rats found exclusively in sub-Saharan Africa (López-Antoñanzas et al., 2004; Woods and Kilpatrick, 2005; Hoffmann, 2008). Indeed, grasscutter and the lesser cane rat (*T. gregorianus*) are the only known extant members of the genus *Thryonomys* and the family Thryonomidae (Woods and Kilpatrick, 2005; Merwe, 2015). Thrynomidae, Petromuridae, and Bathyergidae make up Phiomorpha, one of the early colonization of Hystricognaths and even rodents in African (Huchon and Douzery, 2001; Poux et al., 2006). Fossil evidence suggests that Phiomorpha might have had many members that are now extinct (Bate, 1947; López-Antoñanzas et al., 2004; Kraatz et al., 2013; Sallam and Seiffert, 2016; Sallam and Seiffert, 2020). However, the cane rats, the dassie rats, and the blesmols are probably the only known extant species of Phiomorpha (Huchon and Douzery, 2001; Sallam et al., 2009; Sallam and Seiffert, 2016), suggesting that these species must have evolved strong adaptive traits. Despite the relatively few species in Phiomorpha, the phylogeny is still under debate (D'Elía et al., 2019; Sheng et al., 2020). Among the known extant Phiomorpha species, the cane rats, especially the grasscutter (also called the greater cane rat), has the largest body size (Baptist and Mensah, 1986a). Despite the large body size, grasscutter is a good runner and swimmer; hence, it has a relatively wider geographical distribution (Woods and Kilpatrick, 2005; Hoffmann, 2008) than several other Phiomorpha species. Consequently, cane rats, mainly the grasscutters have been hunted for their meat across many countries in sub-Saharan Africa (Baptist and Mensah, 1986a; Kalu and Aiyeloja, 2002; Annor et al., 2008; Andem, 2012; Coker et al., 2017; Yisau et al., 2019).

One significant physical difference between the grasscutter and the lesser cane rats is their body size. Grasscutter can weigh up to 6 kg (Baptist and Mensah, 1986a; Baptist and Mensah, 1986b; Van Der Merwe, 1999; Merwe, 2015), but the lesser cane rat weighs less. The bigger body weight, unique meat flavor and the relative abundance of grasscutter in West Africa have made the animal an important source of animal proteins for human populations, especially in the rural areas. Grasscutter is an important game animal with desirable meat qualities (Kalu and Aiyeloja, 2002; Yisau et al., 2019; Teye et al., 2020), therefore it has attracted considerable scientific interests (Annor et al., 2008; Owusu et al., 2010; Andem, 2012; Adu et al., 2017; Coker et al., 2017; Yisau et al., 2019; Durowaye et al., 2021). Efforts are now being made to improve its domestication as micro-livestock, while the wild populations are continuously being hunted for human consumption.

Grasscutters are naturally adapted to the reeds and sugar cane farms, but recent human anthropogenic activities have drastically made them adapt to a wide range of habitats, including even urban areas (Wilson and Reeder, 2005; Hoffmann, 2008; Kilwanila et al., 2021). However, their distribution has been somewhat limited to certain parts of sub-Saharan Africa (López-Antoñanzas et al., 2004; Hoffmann, 2008; Coker et al., 2017). They are common animal pests found in grasslands and cultivated forest regions of sub-Saharan Africa (Fayenuwo and Akande, 2002), posing a threat of huge economic loss to the crop farmers. Consequently, in addition to the meat, another motivation for grasscutter hunting is for pest control (Fayenuwo and Akande, 2002; Aluko et al., 2015; Essuman and Duah, 2020) (Fayenuwo and Akande, 2002; Aluko et al., 2015; Essuman and Duah, 2020). Therefore, the animals have a great economic importance both in agriculture and human dietary animal protein supply chain (Adenyo et al., 2012).

Generally, animals with large body sizes tend to have smaller litter size and fewer litter frequency (Tuomi, 1980; Babarinde and Saitou, 2020). The average litter size of grasscutter is 2.9 (Van Der Merwe, 1999; Adu et al., 2000), while the maximum of two litters per female is reported per year (Van Der Merwe, 1999)· The relatively low reproductive rates of this animal and high hunting intensity with no regulations (Van Der Merwe, 1999; Andem, 2012; Yisau et al., 2019; Teye et al., 2020) should suggest grasscutter to be among the threatened or endangered wildlife species (Aluko et al., 2015). However, grasscutter is classified as "Least Concern" animal by the International Union for Conservation of Nature (Child, 2016), implying that the animal does not require any urgent conservation efforts (Hoffmann, 2008; Adenyo et al., 2012).

Previous studies on grasscutter at the population genetics level are scarce and with limited scope. For example, Adenyo et al. (2013) studied exclusively Ghanaian grasscutter population using mitochondrial D-loop region, while other studies have employed microsatellite markers (Adenyo et al., 2017; Coker et al., 2017). Another study on bush meat included grasscutter mitochondrial markers (Gaubert et al., 2015) but did not focus on grasscutter population genetics. It is noteworthy that study on grasscutter population genetics using mitochondrial nucleotide sequences in Nigeria is yet to be documented (Mustapha et al., 2020). Importantly, the impacts and the threat of hunting to the wild grasscutter population has not been extensively investigated.

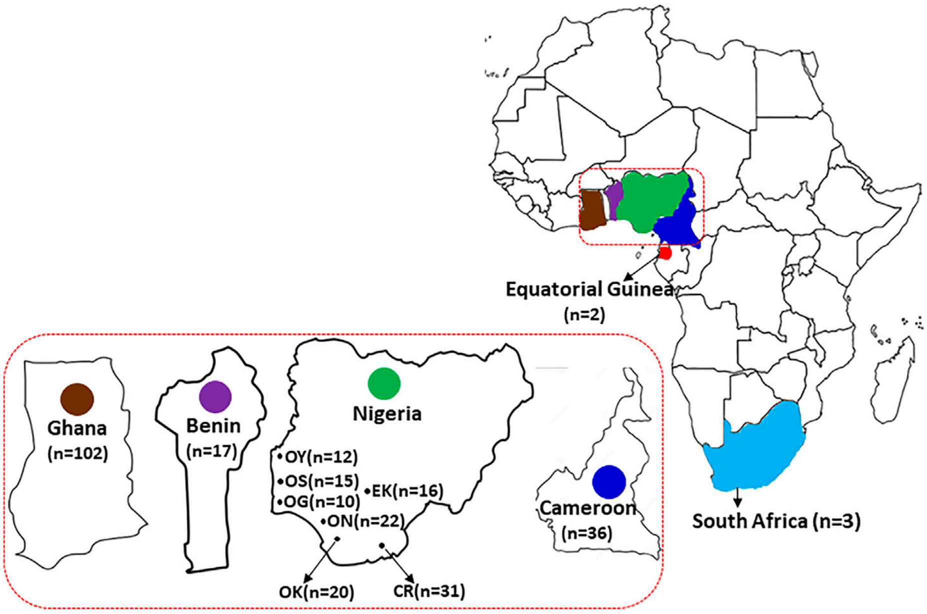


FIGURE 1
The geographical distribution of the grasscutter populations. The samples were collected from countries in the Guinean Forests of West Africa, including Republic of Benin, Nigeria, and Cameroon. Nigerian samples were collected from seven different states. Sequences of samples collected from Ghana, Equatorial Guinea and South Africa were also analyzed.

Focusing on Nigeria, the country with the largest land area in West Africa, the largest human population in Africa and potentially higher threat on wild grasscutter population due to hunting, this study aimed at investigating the wild grasscutter populations in and around Nigeria. The analyses of the demographic histories would reveal any potential threat to the wild grasscutter populations. Also, maximizing breeding gains from grasscutter requires adequate understanding of the genetic diversity and the population structures of the wild populations from where the breeding stock would be selected. Therefore, we analyzed the mitochondrial genome sequences to investigate the population structures and history of wild grasscutter populations in Nigeria and neighboring countries in the African Guinea forests including Republic of Benin, Cameroon, and Ghana. Our results not only help understand the genetic diversity and population structures of Nigerian wild grasscutter populations, but also provide insights into potential evolutionary history of wild grasscutter populations in the African Guinea forests. The findings from the study would provide valuable insights into wild grasscutter breeding stock selection for the purpose of domestication.

2 Materials and methods

2.1 Ethical statement

T. swinderianus is not protected under any legislation and not considered threatened or endangered. Samples from Nigeria were collected through capture and release from National Parks and permissions were collected from the Nigerian National Park Service (NPH/GEN/121/XXV/675). Samples from Republic of Benin (586/DGEFC/DCPRN/SCPRN/SA) and Cameroon were collected from bush meat markets. We have complied with ARRIVE at submission.

2.2 Sample collection

Hair samples plucked to the root were collected from 121 wild grasscutters sampled from seven Nigerian states which belong to three different vegetational zones that support wild grasscutter populations. The sampling locations which are areas with intense grasscutter hunting included Oyo and Ekiti (derived savanna), Osun, Ogun, Ondo and Edo (rain forest) and Cross River (humid rainforest) states (Figure 1). Five of the locations are in the Southwest zone while Edo and Cross River states are in the South-South zones of Nigeria. Additional individuals were sampled from bush meat markets across different vegetational zones in Cameroon (n = 36) and Republic of Benin (n = 17). All samples were collected from January to March 2019 (Supplementary Tables S1). Hair root samples collected were preserved in 95% ethanol and stored under −80°C at the State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, China.

2.3 DNA extraction, PCR amplification and sequencing

Genomic DNA extractions were performed following the standard phenol-chloroform method (Sambrook and Russell, 2001; Akinwole and Babarinde, 2019). The extracted DNA was quantified using the Thermo Scientific™ NanoDrop 2000 spectrophotometer to assess purity. Furthermore, the DNA extracts were checked for molecular quality by running them through a 2% agarose gel together with a 2 kb DNA ladder marker. The five mitochondrial fragments were sequenced in 175 grasscutters samples using primer pairs amplifying 384–658 bp fragments of D-loop, cytochrome b (*CYTB*), cytochrome c oxidase I (*COI*), ribosomal subunits 12 S and 16 S (Supplementary Tables S1, S2). The amplification was performed on a GeneAmp® PCR system 9,700 Applied Biosystems in a 50 µL volume containing PCR mixture of 5 µL 10x reaction buffer, 1.5 mM MgCl2, 0.2 mM dNTPs, 0.2 µM each primer, 1.5 µ Taq DNA polymerase (TaKaRa), and approximately 30 ng genomic DNA. PCR cycling conditions included an initial denaturation of 95°C for 2 min, followed by 35 cycles of 95°C for 30 s, annealing (30 s; see Supplemetary Table S2 for T), an extension of 72°C for 30 s and a final extension of 72°C for 15 min. The PCR products were purified with ExoSAP-IT as per the manufacturer's instructions (Affymetrix). Sequencing reactions were performed using the BigDyeTM Terminator Cycle Sequence Kit 3.1 Ready Reaction Cycle Sequencing Kit (ABI Applied Biosystems), and the products were purified by alcohol precipitation. The purified products were analysed in ABI PRISM 3730 automated DNA sequencer (ABI Applied Biosystems). The electropherograms for each sequence were visualized, edited, and aligned by SeqMan Pro of DNASTAR Lasergen 7.1.0 (DNAStar Inc., Madison, WI) with the reference sequence Accession AJ301644 (Mouchaty et al., 2001).

2.4 Dataset assembly

The identities of the newly generated sequences were confirmed by BLAST searches (Altschul et al., 1997) in the National Center for Biotechnology Information (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The nucleotide sequences of all the newly sequenced samples were deposited in GENBANK under accession numbers MZ418538 - MZ418687; MZ418390 - MZ418537; MZ418252 - MZ418389; MZ418839 - MZ418996; MZ418688 - MZ418838 for D-loop, *CYTB*, *COI*, ribosomal subunits 12 S and 16 S respectively. In addition, previously published D-loop (*n* = 86), *CYTB* (*n* = 25), *COI* (*n* = 26), 12 S ribosomal subunit (*n* = 22) and 16 S (*n* = 27) sequences of *T. swinderianus* from Ghana, Nigeria, Equatorial Guinea, and South Africa (Mouchaty et al., 2001; Gaubert et al., 2015) were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov) (Supplementary Table S1). Further, the nucleotide and amino acid sequences of multiple rodent species were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov). In total, 236 D-loop, 173 *CYTB*, 164 *COI*, 180 ribosomal subunits 12 S and 178 16 S nucleotide sequences of *T. swinderianus* were analysed in this study. The *CYTB* sequences of dassie rat (*Petromus typicus*, accession number DQ139935.1), naked mole rat (*Heterocephalus glaber*, accession number NC_015112.1) and guinea pig (*Cavia porcellus*, accession number NC_000884.1) were also included in the estimation of divergence times.

2.5 Initial data analyses and sequence alignment

First, the sequences of each region were aligned in MEGA7 (Kumar et al., 2016) using CLUSTALX 2.1 (Larkin et al., 2007) with default parameters. For quality assessment, the aligned sequences of *CYTB* and *COI* were independently translated into amino acids using the vertebrate mitochondrial code. No premature stop codons were observed, demonstrating that the open reading frame was maintained in the protein-coding loci. In all the loci, no unexpected gap was found within the alignments.

2.6 Population analyses

2.6.1 Genetic diversity

Sequence comparison and identification of haplotypes were performed with DNASP 5.10.1 (Librado and Rozas, 2009). Genetic diversity was estimated using Arlequin v3.5 (Excoffier and Lischer, 2010) and expressed in terms of number of haplotypes (nHT), haplotype diversity (HTdiv), nucleotide diversity (πdiv), mean number of pairwise differences (MNPD) and their respective standard deviations estimated across all populations used in this study. We ran the analyses on the five concatenated regions for Nigerian samples. Based on the result of the PartitionFinder2, we further analysed the two clusters of mitochondrial regions for Nigeria and cross-countries samples. Because all the five regions were not concurrently sequenced in the same Ghana samples (Supplementary Table S1), all the five mitochondrial regions could not be combined for cross-country analyses.

2.6.2 Phylogenetic analyses

Because of the limited number of variable sites, we explored the possibility of merging all the five regions. We first evaluated the molecular evolution models for the five mitochondrial regions using PartitionFinder2 (Lanfear et al., 2012). The best partition scheme, ranked by the Bayesian information criterion (BIC), separated the regions into two clusters. The first cluster included D-loop while the other four regions were clustered together. The best evolutionary model for the two clusters was HKY (Hasegawa et al., 2005) with invariant site (HKY + I). We therefore analysed the D-loop independently. The other four regions (*CYTB*, *COI*, 12 S and 16 S), which constituted the second partition, were concatenated and further analysed as a unit.

The phylogenetic analyses were conducted using both maximum likelihood (Felsenstein, 1981) and Bayesian inference (Rannala and Yang, 1996; Mau and Newton, 1997) methods. First, the best evolutionary model for each multiple sequence alignment was determined using PartitionFinder2 or MEGA7 (Kumar et al., 2016). The phylogenetic trees were then constructed with the selected best-fit evolutionary models. The maximum likelihood phylogenetic analyses were conducted in MEGA7 with bootstrap test set at 1,000 replications to assess the confidence of the nodes (Felsenstein, 1985). The Bayesian inference trees were constructed with BEAST v2.6.6 (Bouckaert et al., 2019). The priors were set using BICEPS model (Douglas et al., 2021; Bouckaert, 2022). Strict clock with clock rate of 1.0 was used. The MCMC chain length of 50 million was used. Pre-burnin was set to 10% of the total run, while the number of initialization attempts was set to 1,000, with every 1,000 samples being stored. The appropriateness of the MCMC run was evaluated with Tracer v1.7.2 (Rambaut et al., 2018). TreeAnnotator in BEAST package was used to analyze the trees to obtain the tree with the maximum clade credibility based on median heights. To further visualize the genetic relationships between the haplotypes, we constructed a median-joining network (MJ) (Bandelt et al., 1999) using the default setting of weights of both transversions and transitions as implemented in NETWORK 4.6.11 software (http://www.fluxus-engineering.com). When needed, the networks were cleaned using maximum parsimony (MP) options.

2.6.3 Demographic dynamic profiles and population genetic structure parameters

To investigate the demographic patterns and population dynamics of the grasscutter populations, demographic statistical parameters for Tajima's *D* (Tajima, 1989), Fu's *Fs* (Fu, 1997) and Harpending raggedness (Harpending, 1994), and population F_{ST} were calculated using ARLEQUIN v3.5.1.3 (Excoffier and Lischer, 2010). To further investigate the signatures of population structure, the haplotype frequencies of the populations were compared (Raymond and Rousset, 1995). Additionally, population haplotype mismatch distribution patterns were estimated (Rogers and Harpending, 1992). To further infer the genetic variation within populations, among populations, and groups of the grasscutter populations, analysis of molecular variance (AMOVA) was conducted with 50,000 permutations in ARLEQUIN v3.5 software. This analysis was conducted at various hierarchical levels. The significant levels for each hierarchical cluster tested were evaluated using the F_{ST} parameter at a significant *p* level of 0.05.

For the Bayesian inference of population size dynamics, the control file was generated with Beauti and the analysis was carried out in BEAST2 (Bouckaert et al., 2019). Based on the result of PartitionFinder2, HKY evolutionary model (Hasegawa et al., 2005) was used. There was no sufficient fossil and nucleotide sequence data to estimate the substitution rate. We also could not use substitution rate of other rodent species because the analyses of nuclear DNA sequences have revealed heterogeneity in rodent evolutionary rates (Babarinde and Saitou, 2013, 2020). Consequently, we used strict clock with clock rate of 1.0, and left the results in substitution rate units. The substitution rate, proportion of invariant sites, Kappa and frequencies were estimated from the data. The priors were set using default values of the BICEPS model (Douglas et al., 2021; Bouckaert, 2022). The MCMC chain lengths of 10, 50, 75 and 100 million were used, depending on the data. Pre-burnin was set to 10% of the total run, while the number of initialization attempts was set to 1,000, with every 1,000 samples being stored. The appropriateness of the MCMC run was evaluated with Tracer v1.7.2 (Rambaut et al., 2018), and the MCMC chain length was increased if there was need. The trace and the tree files were analysed using Bayesian Skyline Analyses in Tracer. Defaults values were used, except for the maximum time being set to "median".

2.6.4 Relationship between the genetic distance and the geographical distance

Because some of the samples were collected from meat markets, it was difficult to obtain the exact locations of all the samples. Hence, only the approximate locations were used, assuming that the animals sampled at a market were hunted from a location close to the market. We followed the procedure used in a Muscovy duck study (Adeola et al., 2022). For each state or country, we obtained the longitude and latitude of the central location. The coordinates of each pair of the location are then used to infer the geographical distance in kilometers. It is important to stress that this gross approximation of geographical distances would be less accurate in populations that are geographically close. However, we used this approximation as a proxy for the relationship. The relationships between the geographical distance and the genetic distance (F_{ST} values) computed from ARLEQUIN, were then investigated and presented in terms of correlation coefficients and scatter plots. This analysis was done separately for both the four concatenated mitochondrial regions, and the D-loop region.

2.6.5 Genetic component analyses

The analyses of the genetic ancestry were first computed in STRUCTURE version 2.3.4 (Pritchard et al., 2000). We ran the analyses separately for the two partitions. The data were coded such that nucleotides A, C, G and T were coded as "11", "22", "33", and "44", respectively. Because the mitochondrial genome is not diploid, we coded the second allele as "0" and indicated "0" as the missing data in STRUCTURE. Only the polymorphic positions were used. The analyses were run with the Pre-burnin set to 5,000 before 10,000 MCMC replications. Models with or without admixture were tested, with alpha value inferred from the data, starting from 1.0. Independent allele frequency was selected with lambda set to 1. The analyses were run in 10 iterations for k = 2 to 11 for all populations. In addition, model involving migration was also tested. To test various values of k, we first checked the estimated log probability of data. We then focused on the k value with the highest Δk (Evanno et al., 2005). We ran the analysis separately for the four merged regions (*COI*, *CYTB*, 12 S and 16 S) and D-loop.

We further investigated the genetic components with the discriminant analysis of principal components (DAPC) (Jombart et al., 2010) implemented in adegenet package (Jombart, 2008). Focusing on the polymorphic positions, the data were coded such that A, C, G and T were represented as 1, 2, 3 and 4, respectively. The matrix of the positions, with individuals as rows and positions as columns, was then made. The matrix was converted to the DAPC data using *df2genind* in adegenet package. The data were first converted into principal components (PCs). The top 30 *p*Cs were used for the analysis. The PC-transformed data were then used as inputs for DAPC. The results were presented using *compoplot* in adegenet package.

2.6.6 Population phylogenetic tree

The population phylogenetic trees were made for the four concatenated and D-loop regions using POPTREE2 (Takezaki et al., 2010). Corrected F_{ST} values were used for the computation of distances while NJ method (Saitou and Nei, 1987) was used for the phylogenetic reconstruction. Bootstrap test with 1,000 replications was perform to test the reliability of the branches.

2.6.7 Pairwise nucleotide distances and principal component analysis

The pairwise nucleotide distances were computed with MEGA7 using maximum composite likelihood method with 5 Gamma parameters. Bootstrap method with 1,000 replicates was used for the estimation of variance. The principal component analysis (PCA) was computed according to the procedure reported by Babarinde and Saitou (2020). Briefly, pairwise genetic distances estimated from MEGA7 were converted into a full matrix. The *prcomp* in R base (https://www.R-project.org/) was then used to compute the PCA from the pairwise distances.

2.7 Divergence time estimates

Mitochondrial *CYTB* haplotypes of grasscutter samples were used for the estimation of the divergence times. In addition, the *CYTB* sequences of dassie rats (*Petromus typicus*, accession number DQ139935.1), naked mole rat (H*eterocephalus glaber*, accession number NC_015112.1) and guinea pig (*Cavia porcellus*, accession number NC_000884.1) were also retrieved. The best evolutionary model for the aligned sequences was checked. The divergence time estimates were then computed using *BEAST (Heled and Drummond, 2010) implemented in the v2.6.6 of BEAST (Bouckaert et al., 2019). Multispecies coalescent model was used for the estimation (Heled and Drummond, 2010; Barido-Sottani et al., 2018; Zhang et al., 2018). For site model, TN93 (Tamura and Nei, 1993) was used with Gamma Category Count of 5, Shape estimated from the data, starting with 1. Kappa1 and Kappa2 were both estimated starting from 2.0. Empirical frequencies were used. Random local clock model with scaling was used. The clock rate was set to be estimated from the data. For the multispecies coalescent model, the population mean of the species population size was estimated from the data starting from 1.0, with linear population function. The ploidy for Y or mitochondrial was used.

The priors for the multispecies coalescent models were set as follows. Yule model was used for the initial tree. Poisson distribution was assumed for the rate changes. For all the other parameters, log Normal distributions were assumed with the values estimated from the data. The calibration points used include 17.6–28.1 million year divergence [M = 3.14, S = 0.12] for dassie rats and cane rats (Fabre et al., 2012; Patterson and Upham, 2014; Upham and Patterson, 2015), 32.6–39.4 million-year divergence [M = 3.59, S = 0.05] for naked mole rats and cane rats (Patterson and Upham, 2014; Upham and Patterson, 2015) and 41.4–49.5 million-year divergence [M = 3.82, S = 0.04] for guinea pigs and cane rats (Poux et al., 2006; Phillips, 2015; Upham and Patterson, 2015). All the calibration branches were treated as monophyletic with log Normal distributions. The chain length of the MCMC run was 100 million with the tree sampled at every 1,000 runs. Pre-burnin was 5 million. The MCMC analysis was evaluated with the Tracer. The consensus tree was then made by TreeAnnotator in BEAST package. The tree with the maximum clade credibility based on median heights was selected using burnin percentage of 10%. The tree was visualized using FigTree (tree.bio.ed.ac.uk/software/figtree).

3 Results

3.1 Genetic structure of Nigerian grasscutter populations with concatenated mitochondrial regions

To investigate the genetic status of the Nigerian wild grasscutter populations, we first investigated the phylogenetic relationships among the individuals sampled across the investigated locations (Figure 1, Supplementary Tables S1). To maximize the number of informative sites, we assessed the evolutionary models of the five mitochondrial regions. The results of the PartititionFinder2 showed that the five regions could be clustered into two schemes, both having HKY + I as the best model (Supplementary Figure S1A). Although the pairwise genetic distances varied across regions (Supplementary Figure S1B), we decided to concatenate the five mitochondrial regions for the Nigerian populations because all the schemes had the same evolutionary model (Supplementary Figure S1A). The phylogenetic trees made by Bayesian inference (Figure 2) and maximum likelihood (Supplementary Figure S2A) showed that individuals from the same populations were not uniquely clustered. Interestingly, some Cross River samples clustered together with high posterior and bootstrap values. The haplotype network showed that many of the haplotypes of the merged regions have low frequencies (Supplementary Figure S2B). Indeed, very few were found in more than one population. The presence of few population-specific phylogenetic clusters might be attributable to admixture or recent population divergence.

To check the admixture hypothesis, we investigated the population structures of the sampled populations using the five concatenated mitochondrial regions. The overall fixation index for the Nigerian populations was 0.32306. AMOVA (Excoffier et al., 1992) showed that 67.69% of the molecular variance in Nigerian populations was within population, while less than a third of the total variance was among populations (Supplementary Table S3, *p*-value < 10^{-5}). The overall exact test of differentiation based on the haplotypes frequencies (Raymond and Rousset, 1995) was also significant (*p*-value < 10^{-5}). Pairwise comparisons of the F_{ST} values further showed the existence of population structures across various Nigerian population pairs (Supplementary Table S4). Consistent with the observation in the phylogenetic trees, the pairwise comparison between the Cross River population and other Nigerian populations showed significant F_{ST} values. The pairwise exact test of differentiation based on the haplotype frequencies similarly showed significant values across certain Nigerian populations (Supplementary Table S5).

After establishing the existence of some level of population structure across multiple Nigerian populations, we then proceeded to check the demographic dynamics of the populations. The skyline plots showing the population size dynamics over time revealed that the population size remained stable in Ekiti, Osun and Edo populations (Supplementary Figure S3). Ondo and Cross River populations showed slight recent population expansion but the mismatch distribution (Harpending, 1994) showed that the population size changes were not statistically significant in any of the investigated populations (data not shown). It is important to note that the number of sample sizes for the individuals having the five regions were not enough to estimate demographic history in Oyo and Ogun. These data established the existence of a certain degree of population structure in Nigerian grasscutter populations with little evidence for recent population expansion.

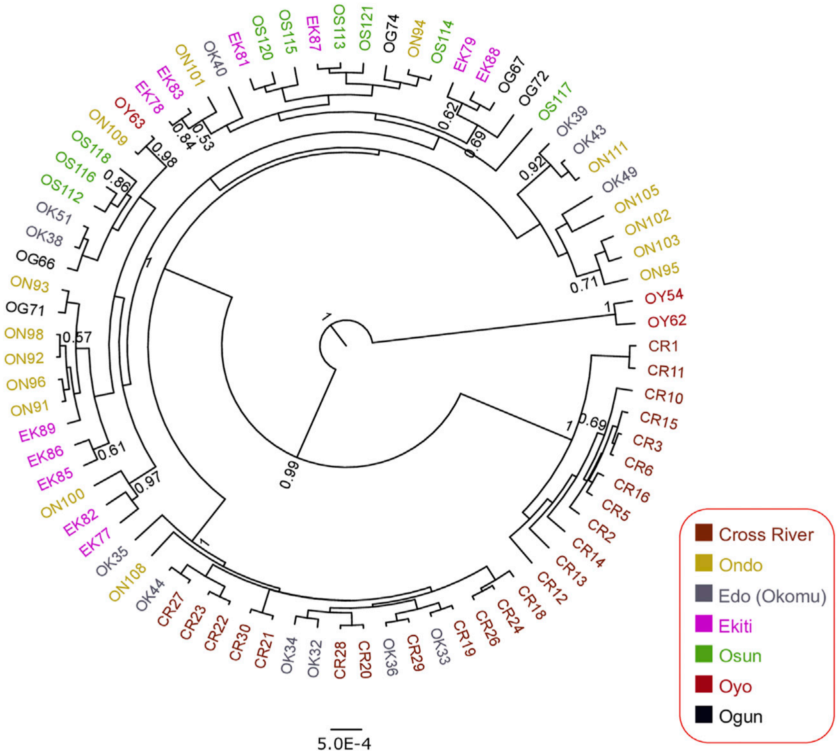


FIGURE 2
Bayesian phylogenetic tree for 79 Nigerian wild grasscutter samples based on the 2,437 bp of the five concatenated mitochondrial regions. The samples are colored by the sampling location. The rooted tree with the maximum clade support from 45,000 MCMC trees is presented. Branches with at least 0.5 posteriors are shown.

3.2 Detailed analyses of the genetic diversity of the Nigerian grasscutter populations with partitioned mitochondrial regions

Having established the existence of genetic structures across numerous Nigerian populations, we then proceeded to investigate the genetic parameters of the populations. Based on the results of PartitionFinder2, the five mitochondrial regions were partitioned into two schemes, with D-loop being separated and the other four regions forming a cluster. Although the two schemes had a similar evolutionary model, D-loop tended to have higher pairwise distances than other regions (Supplementary Figure S1). We therefore analyzed the D-loop region separately. The complete set of four concatenated mitochondrial regions (*CYTB*, *COI*, 16 S and 12 S) were obtained in 83 Nigerian samples spread across the investigated locations (Figure 1, Supplementary Tables S1, S6). The total length of the four concatenated mitochondrial aligned regions was 1,936 bp, out of which 42 sites were variant. The 83 Nigerian samples were assigned into 26 haplotypes (Figure 3A; Table 1; Supplementary Table S6). The overall haplotype diversity for Nigerian samples was (0.845 ± 0.030). Location-based haplotype diversities ranged from 0.542 (± 0.147) in Ondo subpopulation to 1 (± 0.127) in Ogun and Oyo subpopulations (Table 1). The nucleotide diversities ranged from 0.101 (± 0.052) in Cross River to 0.667 (± 0.523) in Oyo, with the overall nucleotide diversity for Nigerian grasscutters being (0.041 ± 0.022). Tajima's and Fu's tests of neutrality showed that the locations with significant values had negative values.

The haplotype network of the concatenated regions showed that the major haplotype observed in about 33% of all the Nigerian samples was not detected in Cross River population (Figure 3A). Interestingly, the second most abundant haplotype, in about 19% of the total samples, was found exclusively in Cross River and Edo populations. In addition, the next two most abundant haplotypes were found exclusively in Cross River samples. The phylogenetic trees with Bayesian inference (Figure 3B) and maximum likelihood method (Supplementary Figure S4) clearly showed the phylogenetic distinctness of some individuals from Cross River and two Oyo samples with high statistical supports. Indeed, F_{ST} values suggest the existence of some level of population structure across Nigerian subpopulations (Table 2). Consistent with the results of F_{ST}, the population structure tests using the haplotype frequencies showed that some Nigerian population pairs differed significantly (Supplementary Table S7). The AMOVA revealed the level of population structure across Nigerian subpopulations (Supplementary Table S8). Specifically, 83.59% of the variance component in Nigerian population was within subpopulations, while only 16.41% was among the subpopulations.

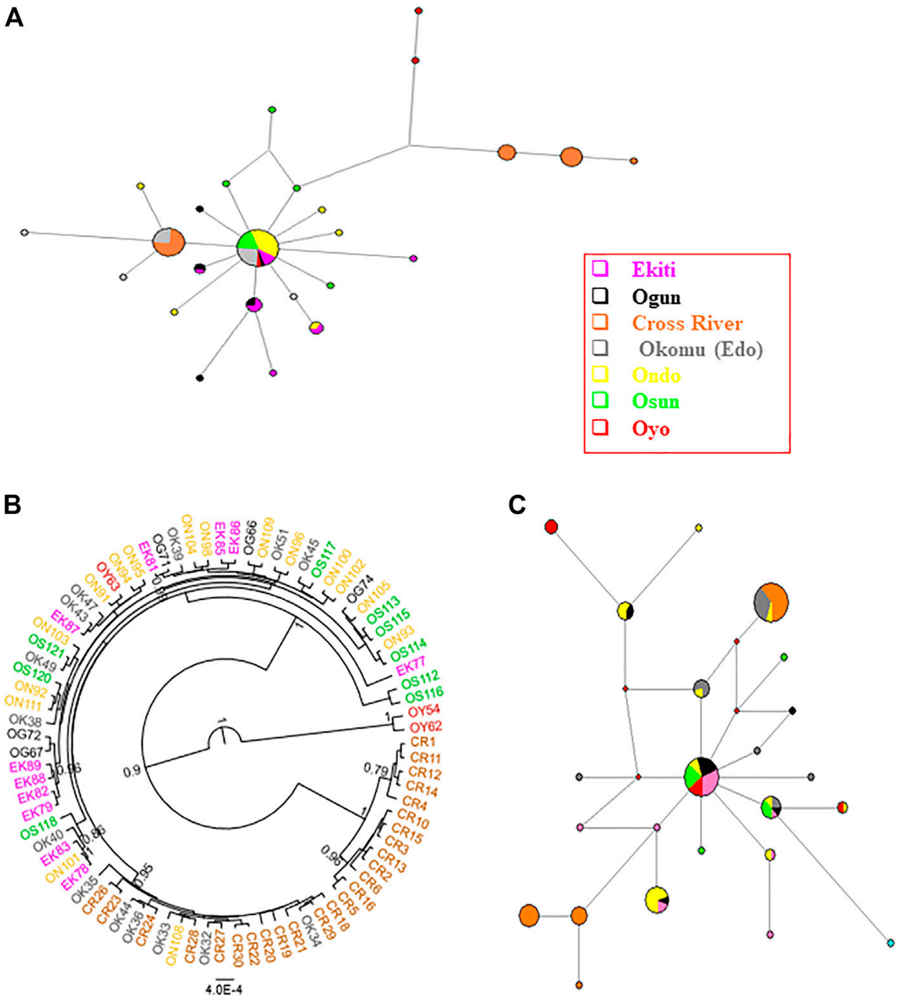


FIGURE 3
Phylogenetic analyses of Nigerian grasscutter population. The haplotypes are colored according to the sampling locations. **(A)** The haplotype network for the Nigerian grasscutters based on concatenated *CYTB*, *COI*, 16 S and 12 S mitochondrial regions. **(B)** The rooted phylogenetic tree computed from Bayesian inference for Nigerian grasscutter sequences based on the concatenated *CYTB*, *COI*, 16 S and 12 S mitochondrial regions. Branches with at least 0.5 posterior probability are shown. **(C)** The haplotype network for the Nigerian grasscutters based on mitochondrial D-loop region.

3.3 Genetic structure of Nigerian grasscutter populations with specific mitochondrial regions

To maximize the number of individuals that could be analysed (Supplementary Table S1), we focussed on the specific mitochondrial regions. Moreover, D-loop region had higher pairwise genetic distances (Supplementary Figure S1) which might be more useful in highlighting recent population history. We therefore first focused on mitochondrial D-loop locus with a larger sample size ($n = 105$; Supplementary Table S1). We identified 35 varying sites, assigned to 22 haplotypes from 478 bp aligned mitochondrial D-loop regions (Supplementary Table S9). The haplotype diversity ranged from 0.659 (± 0.059) in Cross River to (0.837 ± 0.075) in Ondo (Supplementary Table S10). The nucleotide diversity (π) ranged from (0.214 ± 0.127) in Ondo to 0.551 (± 0.335) in Oyo. Although there is a wide range, Tajima *D* and Fu's *Fs* tests for neutrality with D-loop regions did not show significant values for most of the wild populations investigated in Nigeria. The only exception was in Ekiti populations with a significant negative value (−2.672). On the contrary, the analyses of the Tajima *D* and Fu's *Fs* tests on *CYTB* revealed significant negative values for most populations (Supplementary Table S11).

The haplotype network of the Nigerian wild grasscutter populations revealed the distribution of different haplotypes of the D-loop regions (Figure 3C). The two most abundant haplotypes, which represented 43% of the total sampled individuals, had different subpopulation distributions. The most abundant haplotype ($n = 23$ or 22%) was found in Cross River ($n = 14$), Edo ($n = 8$), and to a less extent in Ondo ($n = 1$). The second most abundant haplotype ($n = 22$ or 21%), which seemed to be more

TABLE 1 Genetic diversity of Grasscutter populations based on 1,936 bp concatenated mitochondrial regions.

Population	*N*	nHT	HTdiv	πdiv	*D*	*Fs*	SSD	HRI	MNPD
Cross River	26	5	0.70 (0.06)	0.101 (0.052)	−2.435*	15.742	0.142	0.155	16.892 (7.767)
Ekiti	11	6	0.87 (0.071)	0.259 (0.177)	−0.998	−1.464	0.003	0.042	2.073 (1.254)
Ogun	5	5	1.00 (0.127)	0.450 (0.361)	−0.41	−3.304*	0.052	0.24	1.800 (1.236)
Okomu-Edo	14	5	0.703 (0.101)	0.207 (0.156)	−1.227	−1.092	0.017	0.124	1.242 (0.835)
Ondo	16	6	0.542 (0.147)	0.125 (0.099)	−2.110*	−2.677*	0.002	0.067	1.000 (0.710)
Osun	9	5	0.722 (0.159)	0.306 (0.242)	−0.689	−1.995*	0.012	0.11	1.222 (0.854)
Oyo	3	3	1.000 (0.272)	0.667 (0.523)	0	1.272	0.292	0.667	11.333 (7.123)
Nigeria	84	26	0.845 (0.030)	0.041 (0.022)	−2.703*	−2.992	0.025	0.043	7.743 (3.643)
Benin	11	6	0.873 (0.071)	0.242 (0.150)	−1.286	−0.064	0.215	0.08	3.636 (1.994)
Cameroon	24	5	0.638 (0.061)	0.126 (0.092)	−1.770*	−0.293	0.028	0.184	1.257 (0.822)
Ghana	11	9	0.964 (0.051)	0.189 (0.100)	−2.159*	3.486	0.076	0.197	68.691 (32.135)
South Africa	2	1	–	–	–	–	–	–	–

N = Total number of samples analyzed; nHT, number of haplotypes; HTdiv, Haplotype diversity, πdiv = Nucleotide diversity; *D* = Tajima's *D* test of selective neutrality; *Fs* = Fu's *Fs* test of selective neutrality; SSD, sum of square deviation for mismatch distribution; HRI , Harpending's raggedness index for mismatch distribution; MNPD, Mean number of pairwise differences. The values in braces are the standard deviations. Asterisks indicate statistical significance at 5% level. Benin = Republic of Benin.

TABLE 2 Pair-wise difference F_{ST} between subpopulations in Nigeria, Cameroon, Republic of Benin, and Ghana populations based on 1,936 bp of the four concatenated mitochondrial regions.

	Cross river	Edo	Oyo	Ogun	Ekiti	Ondo	Osun	Cameroon	Benin	Ghana
Cross River	-									
Edo	0.121*	-								
Oyo	0.140*	0.718*	-							
Ogun	0.061	0.181*	0.540*	-						
Ekiti	0.127*	0.154*	0.638*	−0.052	-					
Ondo	0.146*	0.085*	0.742*	0.119*	0.075*	-				
Osun	0.095*	0.151*	0.649*	0.115*	0.102*	0.054*	-			
Cameroon	0.177*	0.797*	0.703*	0.785*	0.766*	0.804*	0.778*	-		
Benin	0.062	0.140*	0.500*	0.053	0.103*	0.097*	0.063	0.634*	-	
Ghana	0.153*	0.207*	−0.063	0.063	0.170*	0.225*	0.137*	0.196*	0.125*	-

Numbers with asterisks are statistically significant at $p < 0.05$.

central, and more likely to have contributed to the radiation of several other haplotypes is found in all populations except Edo and Cross River populations. Although, low-frequency haplotypes were exclusively restricted to certain populations, several haplotypes occurred in multiple populations. Similar analyses with *CYTB,* 16 S, 12 S and *COI* mitochondrial regions, which were known to contain sites under various levels of purifying selections, revealed different patterns (Supplementary Figure S6A–D). For example, while 22 different haplotypes were found on the D-loop region, the *CYTB* and 16 S regions had only eight haplotypes (Supplementary Figures S6A, B), suggesting higher haplotype diversity in D-loop region. Indeed, 12 S had only four haplotypes (Supplementary Figure S6C), while *COI* had 25 haplotypes (Supplementary Figure S6D). Furthermore, single major haplotypes represented a significant percentage of the sampled individuals in the four mitochondrial regions. For example, 101 of the 108 12 S region belonged to the same haplotype (Supplementary Figure S6C). The haplotype networks of D-loop and *CYTB* were consistent with the observations of the neutrality tests of Tajima's *D* and Fu's *Fs* (Supplementary Tables S10, S11). Especially for Tajima's *D*, most Nigerian populations had significantly negative values for *CYTB*, but the values were not significant for D-loop region (Supplementary Table S10).

Overall analysis of molecular variance (AMOVA) of D-loop for the Nigerian wild grasscutter populations revealed that 86.14% of the population variance were within-population variance (p-value < 0.001) (Supplementary Table S12). AMOVA of *CYTB* sequences of Nigerian populations revealed that higher percentage of the variance

(92%) was within-population (p-value < 0.001) (Supplementary Table S13). The pairwise fixation index (F_{ST}) computed from D-loop sequences showed consistency with the geographical locations; Cross River population seemed to be isolated from other populations (Supplementary Table S14). Indeed, most of the population pairs had significant F_{ST} values. However, the pairwise F_{ST} values between Ekiti, Ondo and Ogun populations were not significant. Likewise, the F_{ST} value for Osun and Ogun populations was not significant. The tests of significance of F_{ST} showed that Cross River and Oyo subpopulations were genetically isolated from other wild Nigerian grasscutter subpopulations. Similar results were found when the haplotype frequencies of the populations were compared in a pairwise manner (Supplementary Table S15).

3.4 Genetic structure of grasscutter populations across Guinean Forests of West Africa based on the concatenated mitochondrial regions

Having highlighted some of the features of the Nigerian wild grasscutter populations, we then investigated how these features differ across the neighbouring countries in the lower Guinea forests of West Africa. Geographically, Cameroon is close to Cross River (Nigeria), and Republic of Benin is very close to Ogun and Oyo states (Nigeria) (Figure 1). We also retrieved publicly available data from Ghana and South Africa wild grasscutter populations. We first confirmed, with the pairwise distance PCA, that all the samples were of reliable quality. The PCA from the pairwise distances showed that TswiT996 was questionable (Supplementary Figures S7A, B). We therefore excluded the sample from further analysis. Because no individual from Ghana had all the five mitochondrial regions (Supplementary Table S1), we followed the data partition scheme and separately analysed the concatenated four regions (*COI*, *CYTB*, 16 S and 12 S). Note that the D-loop region was analyzed separately. The aligned sequences of the four concatenated regions were 1,974 bp long and included 131 grasscutter samples from Nigeria, Cameroon, Republic of Benin, Ghana, and South Africa. These aligned sequences with 276 variant sites were assigned to 45 haplotypes (Supplementary Table S6). The haplotype diversity ranged from 0.638 (± 0.061) in Cameroon to 0.964 (± 0.051) in Ghana (Table 1). The nucleotide diversity ranged from 0.041 (± 0.022) in Nigeria to 0.242 (± 0.150) in Republic of Benin. Whereas the Fu’s statistic was not significant, Tajima D was significantly negative in Nigeria, Cameroon, and Ghana grasscutter populations. Indeed, the skyline plots support the possibility of recent population expansions in Nigeria and Cameroon, but not for the Ghana population (Supplementary Figure S8). However, Harpending’s raggedness index and sum of square deviation (SDD) were not significant (Table 1).

The network based on the four concatenated mitochondrial regions revealed that many of the haplotypes were country-specific (Figure 4A; Supplementary Table S6). Only two haplotypes were shared between countries. One of the haplotypes (n = 31) was shared between Nigerian and Republic of Benin populations, while the other (n = 3) was shared between Ghana and Republic of Benin samples. This suggested the existence of population structures across countries. The phylogenetic analyses using Bayesian inference (Figure 4B) and maximum likelihood (Supplementary Figure S9) revealed that haplotypes from the same country tended to cluster together. Indeed, pairwise F_{ST} was significant across each pair of countries (Table 2). The only exception was between Nigeria and Republic of Benin where the F_{ST} was not significant. Generally, the geographical distance between the populations correlated (Pearson’s r = 0.58, p-value = 3.3e-05) with the genetic distance between the populations (Supplementary Figure S10A). The test of population structure using the haplotype frequencies (Supplementary Table S7) also showed similar patterns to the results of F_{ST} (Table 2). Again, the haplotype frequencies of some Nigerian subpopulations and population from the Republic of Benin were not significantly different. AMOVA for the concatenated sequences revealed the status of the population structure (Supplementary Table S8). For all the samples from the investigated populations, about 73% of the overall total variation was among population while about 27% was within population.

Population phylogenetic tree using the F_{ST} values computed from the four mitochondrial regions reveal the relationships among the populations. For clearer understanding of the phylogenetic relationship, each of the seven Nigerian populations was treated individually. The result showed that the South Africa population diverged first from the other populations (Figure 4C). Surprisingly, Oyo population was found to cluster with both Ghana and Cameroon populations. Consistent with the network in Figure 4A, Benin population was found to cluster with other Nigerian populations.

To better understand the population phylogenetic tree, we investigated the genetic components of each of the populations. We investigated various numbers of ancestral genetic components (k = 2–11) for models with or without admixture, and model involving migration (Supplementary Figure S11). At all the investigated k values in models without migration, the South African components were not found in any other individuals, consistent with the results of the phylogenetic analyses. However, for the model involving migration at k = 2, some Ghana and Oyo individuals showed some levels of South African components. Interestingly, at k = 3 and above, the genetic components did not substantially change in the models without migration. However, some differences were observed across different k values in the model involving migration. Another obvious difference across the models was the number of admixture events. More admixture events were found in the model involving migrations, while the model without admixture expectedly had the least. Interestingly, admixture was found for some individuals from the Republic of Benin at k > 2, even for the model without admixture (Supplementary Figure S11).

In any case, we proceeded to investigate the “best” k value for the grasscutter populations. For all the k values, the STRUCTURE’s choice criterion was higher for models without migration (Supplementary Figure S12A). The models without migration reached the plateau at k = 3, while the migration model had a peak at k = 4. The model choice method involving ΔK (Evanno et al., 2005) confirmed these k values (Supplementary Figure S12B). Figure 4D showed the distribution of ancestral components in each of the populations at k = 3 for the model without admixture and migration. As expected, South African component was not found in any other population. Consistent with the population phylogenetic tree, the dominant component in Nigerian populations was the minor component in both Ghana

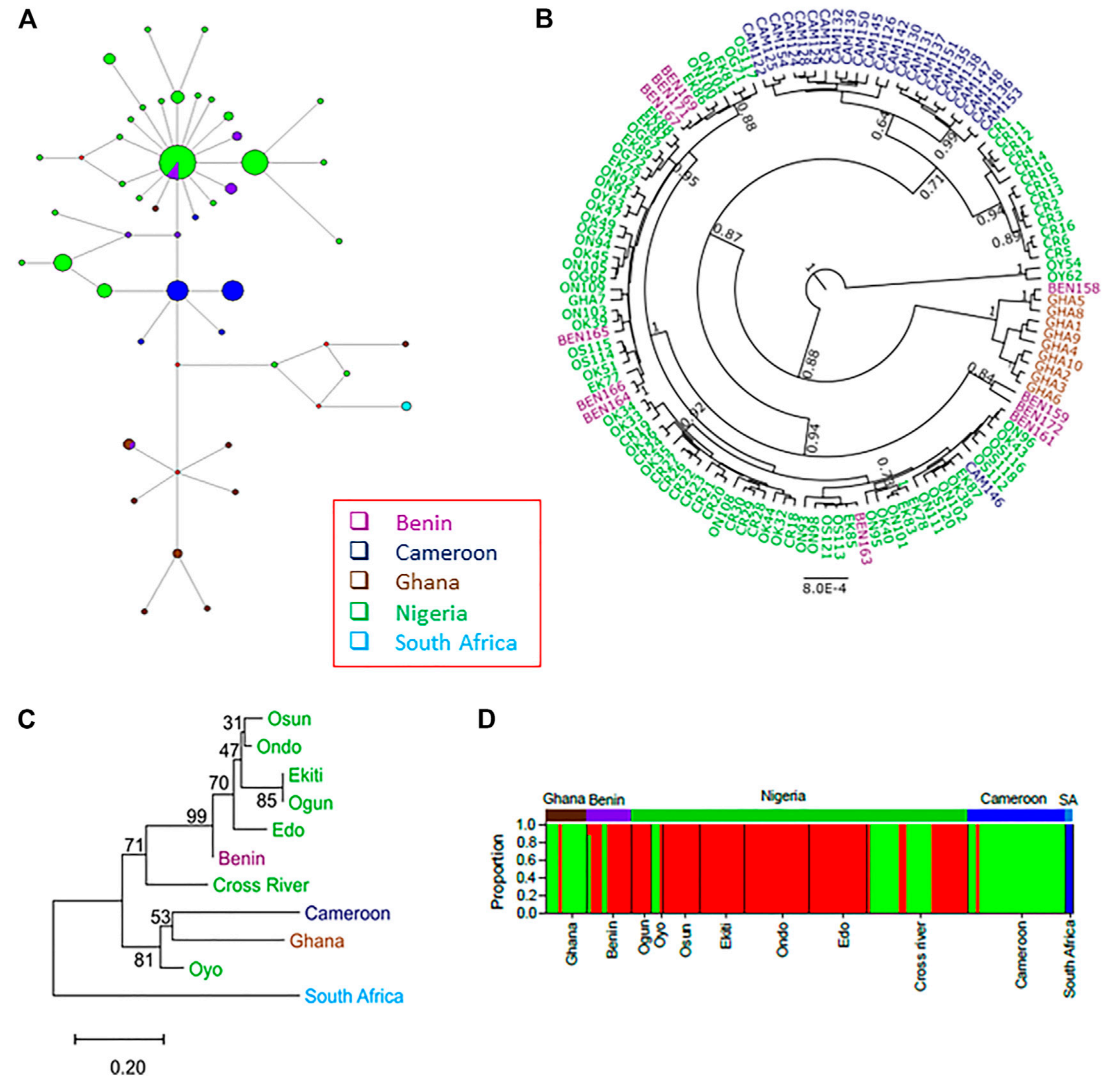


FIGURE 4
The population structures of the wild grasscutters from the Guinean Forests of West Africa based on the concatenated *CYTB*, *COI*, 16 S and 12 S mitochondrial regions. **(A)** The haplotype network for the sampled individuals. **(B)** Phylogenetic tree computed from the Bayesian inference of the grasscutter sequences. Branches with at least 0.5 posterior probability are shown. **(C)** The population tree computed from the corrected F_{ST} values. The bootstrap values from 1,000 replications are shown. **(D)** Structure analysis to reveal the genetic components of various grasscutter populations. The optimum number of component (k) was three. Panels **(A–D)** were made from the concatenated *CYTB*, *COI*, 16 S and 12 S mitochondrial sequences.

and Cameroon populations. Oyo and Cross River populations had some levels of Ghana and Cameroon component. Although the majority of the Benin population had the Nigeria component, some individuals had Ghana and Cameroon components.

Because of the limited number of loci and STRUCTURE's assumptions, we used DAPC to further investigate the genetic components in the populations studied. Consistent with the cluster results for models without migration, South African samples were separated from other populations at all k values tested (Supplementary Figure S13). The results for k = 2 and k = 3 were very similar for the DAPC results and the STRUCTURE's models without migration (Supplementary Figures S11, S13). The genetic components observed for Oyo, Cross River, and Benin in DAPC were similar to the results in STRUCTURE without migration. At higher k values, DAPC was able to capture higher levels of differentiation. At all the k values investigated, samples from the same location tended to have more similar genetic components. Also, samples from the Republic of Benin had similar genetic components to some Nigerian samples. Unlike the results of STRUCTURE, especially when migration model was considered, DAPC did not find multiple signals of admixture.

3.5 Analyses of specific regions reveal more detailed dynamics of grasscutter populations across Guinean Forests of West Africa

We next analysed the D-loop sequences from the countries. A total of 234 analysed grasscutter mitochondrial D-loop sequences

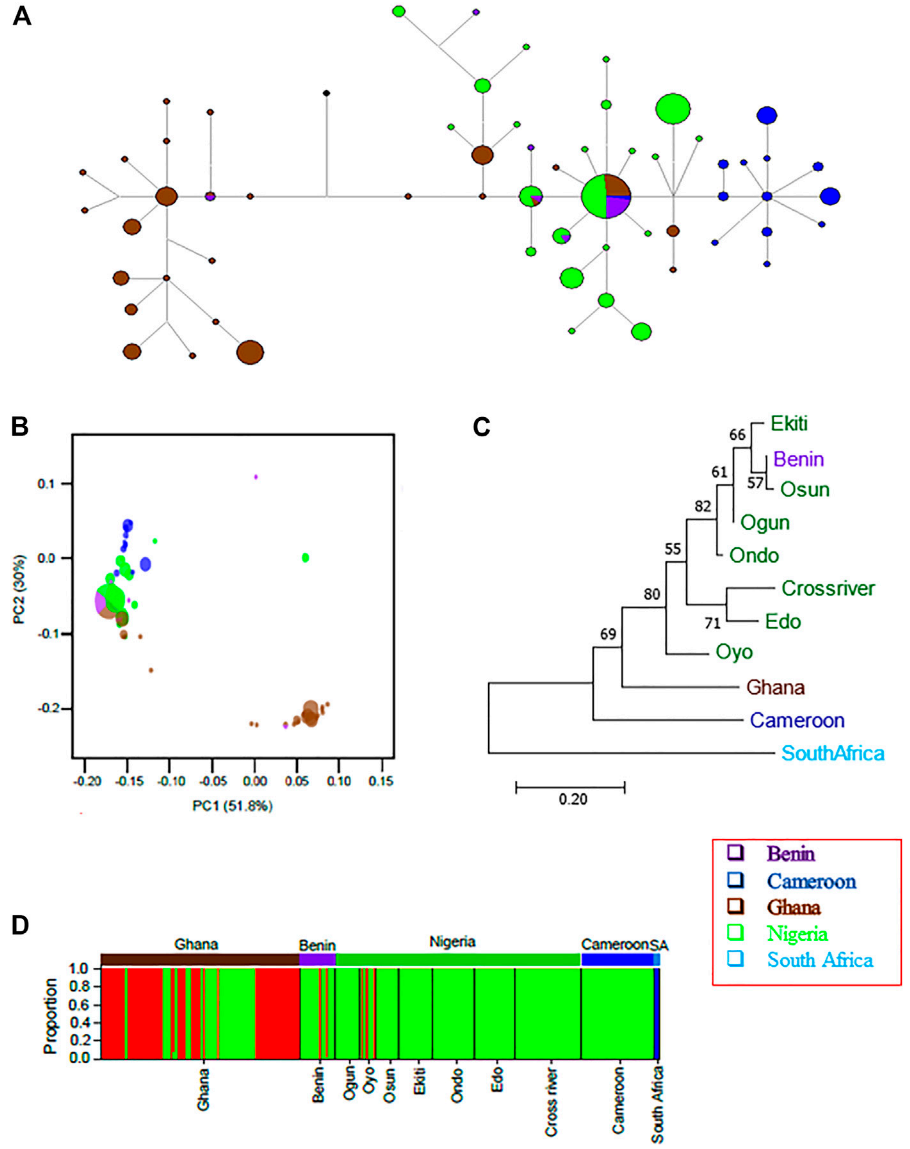


FIGURE 5
Mitochondrial D-loop regions reveal detailed population structure of wild grasscutter populations from Guinean Forests of West Africa. **(A)** The haplotype network for the grasscutters based on mitochondrial D-loop region. **(B)** PCA computed from pairwise genetic distances showing the relationships between various populations. D-loop haplotypes were used for the computation. The colors correspond to the country of sampling while the size is proportional to the haplotype frequency. **(C)** The population phylogenetic tree based on the corrected F_{ST} computed from the D-loop regions. **(D)** Structure analysis to reveal the genetic components of various grasscutter populations. The optimum number of component (k) was three. Panels **(A–D)** were made from the mitochondrial D-loop sequences.

from Nigeria (*n* = 104), Republic of Benin (*n* = 15), Cameroon (*n* = 31) and Ghana (*n* = 84) were assigned to 60 haplotypes (Supplementary Table S9). Unlike in the concatenated regions, the lowest haplotype diversity of D-loop region was found in the Republic of Benin population (0.571 ± 0.149), while the highest in the Ghana population (0.921 ± 0.013) (Supplementary Table S10). Nigerian, Republic of Benin, and Cameroon populations showed statistically significant negative Tajima's *D* values, suggesting population expansions. Indeed, the skyline plots showed evidence of slight recent population expansions in all the populations but the pattern in Ghana was strange with wide confidence interval in the recent years (Supplementary Figure S8). In addition, analyses of *CYTB* revealed similarly lower haplotype and nucleotide diversity in Republic of Benin populations (Supplementary Table S11). However, only Nigerian and Republic of Benin populations showed significantly negative Tajima's *D* values for the *CYTB* region.

Although there were some exceptions, the haplotype network for the D-loop regions closely mirrored the geographical locations of the haplotypes (Figure 5A), and was largely consistent with the concatenated mitochondrial region result (Figures 4A, B; Supplementary Figure S9). The most abundant D-loop haplotype (*n* = 45 or 19%) was found in the four countries under investigation. However, many of the haplotypes in the D-loop regions were specific to each country. The analyses of *CYTB*, *COI*, 12 S and 16 S regions revealed patterns expected under purifying selections or selective sweep with fewer haplotypes dominated by numerous low-frequency haplotypes (Supplementary Figures S14A–D), largely reflecting the results of Nigerian subpopulations. Interestingly, the *CYTB* haplotype with the highest frequency in Nigeria populations (68%), Republic of Benin population (93%) and Cameroon population (55%) are not found in Ghana (Supplementary Figure S14A), suggesting the isolation of Ghana populations.

Despite the haplotype sharing between Republic of Benin and Nigerian populations, the D-loop network showed clear pattern of population structure (Figure 5A), and the AMOVA revealed that 58.98% of the variance was within populations (Supplementary Table S12). This level of within-population variance was lower than the value for Nigerian populations. Consistent with the haplotype network, significant genetic structure was found in all pairs of the populations across countries (Table 2, Supplementary Figure S15). As expected from the geographical distribution, the highest F_{ST} value (0.516) was found between Ghana and Cameroon populations, while the lowest F_{ST} value (0.042) was found between Nigerian and Republic of Benin populations. Indeed, the correlation between the geographical distance and the genetic distance was higher in D-loop region (Person's r = 0.75, *p*-value = 4.9e-09) than in the merged regions (Supplementary Figure S10B). The pairwise tests using the haplotype frequencies (Supplementary Tables S15) produced similar results to F_{ST} tests.

Principal component analyses were then computed using pairwise genetic distances of the D-loop haplotype sequences. PC1, which accounted for about 52% of the variance separate a cluster of Ghanian haplotypes from others (Figure 5B). Also, Cameroonian population seemed to form a cluster that was not too far separated from the Nigerian and Republic of Benin populations, suggesting more recent shared ancestry. To better visualize the relationship, we computed the population phylogenetic tree using the D-loop sequences (Figure 5C). The result confirmed that South African population was the outgroup. Among the other populations, Cameroon population diverged first before the Ghana population diverged from the Nigerian population. Again, Benin population was found to cluster with Nigerian populations. Pairwise F_{ST} analyses using *CYTB* (Supplementary Table S16) revealed essentially similar patterns as the D-loop (Supplementary Table S14). Despite the small sample sizes for South African (n = 3) and Equatorial Guinea (n = 2) populations, the signatures of populations structures were still revealed.

We next checked the ancestral genetic components for the populations based on the D-loop sequences. We checked the components for k = 2 to k = 11 using three different models in STRUCTURE software (Supplementary Figure S16). Consistent with phylogenetic results and the results of the four merged regions, for all the investigated k values, South African samples were consistently separated from the other populations and the genetic components of most of the individuals from the Republic of Benin were found in some Nigerian individuals. Like in the merged regions, more admixture events were found in the model involving migration. Also, the STRUCTURE results were mostly similar for k = 3 and above in both admixture model and the model without admixture. The model involving migration revealed more structures at higher k values. Both STRUCTURE choice criterion (Supplementary Figure S12C) and ΔK (Supplementary Figure S12D) showed that k = 3 best fit our data for models with or without admixture. The best k value for the model involving migration was difficult to resolve (Supplementary Figure S12D). At k = 3 (Figure 5D), Cameroon and Benin components looked more similar to Nigerian component while the Ghana population was more heterogenous. Some individuals in Oyo population also contained different genetic component.

We next repeated the analyses of the D-loop region using DAPC. Like for the STRUCTURE results, the South African samples were consistently shown to be genetically different at all k values investigated (Supplementary Figure S13). The results of DAPC for k = 2 and k = 3 mostly reflected the results of STRUCTURE analyses for models without migration. At higher k values, DAPC revealed more population structures. In all the k values investigated, the population from Benin Republic had more similar genetic components to some Nigerian individuals. At k = 4, four components from South Africa, Cameroon, Nigeria/Benin and Ghana populations were seen. However, at all k values investigated, the analyses of the D-loop showed that some individuals from Ghana have Nigeria/Benin genetic components.

3.6 Phylogenetic analysis and evolutionary history of grasscutter populations based on *CYTB* region

Apart from lesser cane rat, which is in the same genus as grasscutter, the next closest species was dassie rat (*P. typicus*). However, there are no nucleotide sequences of any of the five studied mitochondrial regions for lesser cane rats, and there was paucity of the sequences of mitochondrial regions for dassie rats. Thus, we restricted the divergence time estimate to *CYTB* region with dassie rat sequence (Visser et al., 2019). The *CYTB* sequences formed 13 haplotypes across the investigated countries (Supplementary Table S17). The phylogenetic analyses revealed that the haplotype exclusively found in South African was separated from other grasscutter *CYTB* haplotypes (Figure 6; Supplementary Figure S12A). The South African *CYTB* haplotype was estimated to have diverged from other haplotypes at about 6.07 (2.6–10.18, 95% CI) MYA. Although only three samples from South African were analysed, the fact that the haplotype was observed in none of the 234 West African and Cameroonian individuals (Figure 6), despite the high level of *CYTB* haplotype sharing among West African and Cameroonian populations (Supplementary Figure S12A), suggested an ancient separation of South African population from the other studied populations.

The haplotypes from Nigeria, Cameroon, Republic of Benin, and Ghana shared a much more recent history with the last common ancestor diverging about 0.68 (0.23–1.341, 95% CI) MYA. The phylogenetic relationship of the *CYTB* haplotypes revealed that after the divergence of the South African haplotype, the next haplotypes to diverge were found in Ghana and Republic of Benin. Importantly, one of the Ghanaian haplotypes was also shared by the Republic of Benin population. The sharing of *CYTB* haplotype between Equatorial Guinean and Cameroonian

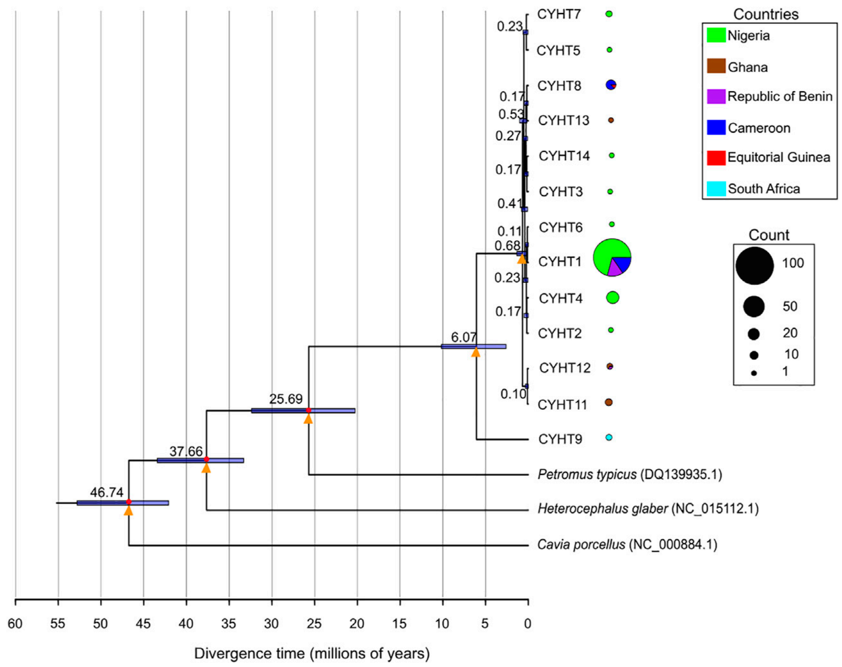


FIGURE 6
Divergence time estimates of the *CYTB* haplotypes. The divergence time estimates are shown in each node. The bars represent 95% of the estimates for the nodes. The calibration points are marked with red circles. The yellow triangles represent nodes with very high statistical supports (posterior probability = 1, bootstrap support value from maximum likelihood method with 1,000 replicate > 95%). The distribution of each haplotype is presented in pie chart, with the size proportional to the total number of the haplotype.

populations was also worthy of note. The phylogenetic analyses revealed that relatively recent emergence of the major *CYTB* haplotype. Because *CYTB* experienced purifying selection, it was expected that many shared haplotypes would still be present even after a long evolutionary period. The relaxed evolutionary pressure in D-loop regions would lead to gradual decrease of shared haplotypes. Taken together, the mitochondrial D-loop and *CYTB* regions presented a more comprehensive picture of the population structure and migration history of wild grasscutter populations across lower Guinean forests.

4 Discussion

This study investigates the population dynamics of Nigerian wild grasscutter population with populations from other African Guinea forest countries using the nucleotide sequences from five mitochondrial regions. While the *CYTB,* 16 S, 12 S and *CO1* are likely to show incomplete lineage sorting, D-loop tends to reveal more recent evolutionary history. Consistent with functional importance, the D-loop region reveals more population dynamics and history, potentially because of the relaxed purifying selection. The overall D-loop haplotype diversity for Nigerian grasscutter population (0.912 ± 0.015) is slightly lower than 1.000 (± 0.016) reported for Nigerian cattle (Mauki et al., 2021), but slightly higher than 0.899 (± 0.148) for sheep (Agaviezor et al., 2012), higher than 0.693 (± 0.022) reported for helmeted guinea fowl (Adeola et al., 2015) and 0.673 (± 0.002) for Nigerian local chicken (Lasagna et al., 2020), suggesting the domestication in the species might have lowered the genetic diversity as previously reported in horses and dogs (Wang et al., 2013; Wang et al., 2015; Fages et al., 2019). Importantly, the high genetic diversity of the wild grasscutter populations suggests that there is no loss of genetic resources. Both haplotype and nucleotide diversities revealed differences across the investigated populations from the African Guinea forests. This information could be useful in both conservation efforts and breeding programs.

The analyses of demographic histories reveals the recent population dynamics of the wild grasscutter populations. Whereas significant values of Fu's *Fs* and Tajima's *D* could reveal recent population size change, they are originally tests for neutrality which could reflect signatures of selections and/or selective sweeps (Tajima, 1989; Harpending, 1994; Fu, 1997). This suggests that the significant values for these parameters might not necessarily reflect population size changes. Indeed, BICEPS results show that the time to the most recent common ancestor (in substitution unit) tends to be different between the four merged regions and the D-loop, reflecting the impacts of evolutionary constraints. As different regions tend to have different evolutionary constraints different degrees of isolation revealed by different markers might reflect deviation from neutrality. Further, the two parameters rely on infinite site model which might not be adequate for mitochondrial genome. Analyses such as mismatch distribution (Harpending, 1994) and Bayesian approach also establish population changes. It is important to point out that BICEPS (Bouckaert, 2022) assumes random mating with no admixture. Therefore, some subtle signals might be missing because of the nature of our data and the colony structures of grasscutters

(Hoffmann, 2008; Coker et al., 2017). Notwithstanding, all the analyses reveal isolated instances of recent population expansion. No signature of population contraction exists, indicating that the wild grasscutter populations are not under threat of extinction. This highlights the classification of the species as "least-threatened" despite the high intensity of hunting.

The determination of the exact number of genetic components in a population or group of populations remains a Herculean task (Pritchard et al., 2000; Evanno et al., 2005). We use two different methods to investigate the best k value under different models in STRUCTURE software. Also, because of the limited number of loci and high possibility that our data might violate some of the assumptions of STRUCTURE, we confirm the STRUCTURE results with DAPC. While the results are essentially similar at lower value, DAPC reveals more structure at higher k values, reflecting the limitation of STRUCTURE models at higher k values, especially when no migration was considered. Whether the additional components detected at higher k values for DAPC and the model involving migration in STRUCTURE are actually genuine or noisy signals could not be ascertained. However, both STRUCTURE and DAPC support our main conclusions.

Although there is high within-population variance and Nigerian populations tend to have the same genetic component, our analyses reveal some level of genetic structures among the Nigerian populations. This is more pronounced when populations are compared across countries. Shared haplotypes are few, especially at the mostly neutrally evolving D-loop locus. The strength of genetic structure relates to geographical distance. The high within-population variance could reflect recent population divergence or admixture, and thus geographical distance. Indeed, a significant positive relationship occurs between the genetic and geographical distances of the populations. However, our data cannot delineate between the recent divergence or admixture hypotheses. Moreover, the genetic distinctness of Cross River subpopulations among the Nigerian grasscutters may be attributable to the elevation of Cross River state, thereby limiting migration. However, the population structure might not be entirely due to migratory limitations as grasscutter have relatively good migratory ability as they can run on land and swim across rivers. One factor that might contribute to the population structure is the colony system (Coker et al., 2017; Mustapha et al., 2020; Kilwanila et al., 2021). Each grasscutter colony comprises a male and several females, thereby creating a form of isolation or structure (Hoffmann, 2008). Owusu et al. (2010) reported decreasing male proportion as litter size increased for grasscutters farms in Ghana. A study of bush meats in Ogun State, Nigeria (Yisau et al., 2019) showed that only 24% of captured juvenile and 40% of sub-adult grasscutter were males, suggesting that the wild litter size ratio might be biased towards females. However, about 61% of captured adults were males. Since most colonies have fewer adult males than females, the colony survival after the death of the breeding male would depend on the taking over by another male from another colony or the replacement by a young male. The replacement by a younger male from the colony would lead to stronger signatures of genetic structure based on maternally inherited mitochondrial DNA.

Although grasscutter is believed to have evolved in Africa (Baptist and Mensah, 1986b; López-Antoñanzas et al., 2004; Hoffmann, 2008), the exact location of the first emergence is not known, and the grasscutter studies have been reported to be biased (Kilwanila et al., 2021). Our results indicate that South African and West African *CYTB* haplotypes diverged at least 6.1 MYA. It is possible that the haplotype divergence may predate population divergence (Maddison, 1997; Suh et al., 2015; Shen et al., 2017). Indeed, the emergence of crown Thryonomidae (the common ancestor of cane rats) has been contested (Kraatz et al., 2013; Sallam and Seiffert, 2016; Sallam and Seiffert, 2020). The inconsistent fossil records and the limited nucleotide sequences for the closely related species restrict fossil calibration to the outgroup species. This could affect the Bayesian estimation of the divergence times. However, the South African samples do not cluster phylogenetically with the samples from other regions in all the investigated genomic regions, suggesting that the divergence is ancient. Although the timing cannot be established with high certainty, the analyses of *CYTB* sequence suggest that Ghana population diverged before the Cameroon population diverged from the Nigerian and Republic of Benin populations. However, the population divergence tree based on D-loop suggests that the Cameroon population diverged first before the split of Ghana and Nigerian/Benin populations. This suggests that *CYTB* might reflect more of gene tree than the population tree. Further, the impacts of purifying selection on the divergence time estimation cannot be fully ascertained. The analyses of nuclear data of different mammalian orders reveal that the use of sites under purifying selection gave more consistent results (Babarinde and Saitou, 2020). However, whether this holds in mitochondrial genomes, and especially in recently diverged population-level individuals is not clear. In any case, our data consistently show that Nigerian and Republic of Benin populations share much more recent history, and that South African population shares a very deep coalescence with other populations.

Grasscutter has been described as a potential livestock for the future (Opara, 2010; Adu et al., 2017). Several efforts are being made in domesticating grasscutters to produce meat for humans (Jori et al., 1995; Adu et al., 1999; Adu et al., 2000; Fayenuwo and Akande, 2002; Adu et al., 2017). Indeed, the meat of the grasscutter is well accepted in West Africa (Aluko et al., 2015; Gaubert et al., 2015; Teye et al., 2020). A classical method of improvement in animal breeding is by heterosis (Birchler et al., 2006; Timberlake, 2013; Akeberegn et al., 2019). Breeding individuals from similar genetic backgrounds can lead to inbreeding depression (Kardos et al., 2016; Hajduk et al., 2018; Kardos et al., 2018; Harrisson et al., 2019). Our study reveals that South African grasscutter crossbred with grasscutters from West Africa would benefit more from heterosis because of their different genetic backgrounds. On the other hand, Republic of Benin and Nigerian populations are not so different genetically. Further, the Republic of Benin population seems to have lower genetic diversity, probably because of the small size of the country. Gain from heterosis with this population would be minimal. Therefore, grasscutter breeding stocks from populations in other countries, outside Republic of Benin, are recommended for grasscutter breed improvement. The second implication of this study is in the understanding of the status of grasscutter populations. This study confirms that the grasscutter populations are generally not threatened.

Although this study reveals some important status of the wild grasscutter populations from the Guinea Forests of West Africa, there are certain limitations of the study. First, the number of individuals that could be analysed is few in some populations. Indeed, the description of the wild grasscutter populations from Equatorial Guinea and South African cannot be extensively investigated because of the extremely low sample sizes. How representative these sampled individuals are for their respective populations cannot be ascertained. Second, the number of informative sites was small. Although the effects of these two factors could be minimized by the number of replications or samplings, the results are still not completely free from stochasticity. Therefore, further studies involving larger sample sizes are recommended for these populations. Finally, this study uses maternally inherited mitochondrial regions with limited number of informative sites, and some analyses that assume infinite site models might not be reliable. Future studies to further confirm the results should be based on nuclear regions sampled from sufficiently large number of individuals from across the locations.

Data availability statement

The data presented in the study are deposited in the NCBI repository with accession numbers MZ418538-MZ418687; MZ418390-MZ418537; MZ418252-MZ418389; MZ418839-MZ418996; MZ418688-MZ418838 for D-loop, cytochrome b (*CYTB*), cytochrome c oxidase I (*COI*), ribosomal subunits 12S and 16S, respectively.

Ethics statement

The animal study was reviewed and approved by National Parks and permissions were collected from the Nigerian National Park Service (NPH/GEN/121/XXV/675). Samples from Republic of Benin (586/DGEFC/DCPRN/SCPRN/SA) and Cameroon were collected from bush meat markets. We have complied with ARRIVE at submission.

Author contributions

ACA, RWM, IAB, LMN and CAMSD designed the study; AOO, OCO, OO, CAMSD, NG, WKN, LMN, BEA, OO, OCO, MMM, AOA, OJS and VMOO collected the samples; ACA, LMN and YYW performed the molecular laboratory work and generated the sequence data; ACA, SIN and IAB performed the genetic analyses; LMN provided technical assistance for the study; IAB and ACA wrote the initial draft of the manuscript; RWM, SOS, CDN, LMN, AOA, CAMSD, WAO, AOO, VMOO, SFB, WKN critically revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Sino-Africa Joint Research Center, the Chinese Academy of Sciences (SAJC202103), the Animal Branch of the Germplasm Bank of Wild Species, and the Chinese Academy of Sciences (the Large Research Infrastructure Funding). The Chinese Academy of Sciences President’s International Fellowship Initiative provided support to ACA (2021FYB0006).

Acknowledgments

We thank the management of Nigerian National Park Service for the permits and provision of technical assistance for the field survey. We also thank individuals that assisted in collecting samples from bush meat markets for this study.

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 reviewer LHD declared a shared affiliation with the author CAMSD to the handling editor at the time of review.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1041103/full#supplementary-material

References

Adenyo, C., Hayano, A., Inoue, E., Kayang, B. B., and Inoue-Murayama, M. (2012). Development of microsatellite markers for grasscutter (Thryonomys swinderianus, RODENTIA) using next-generation sequencing technology. *Conserv. Genet. Resour.* 4, 1011–1014. doi:10.1007/s12686-012-9695-5

Adenyo, C., Hayano, A., Kayang, B. B., Owusu, E. H., and Inoue-Murayama, M. (2013). Mitochondrial D-loop diversity of grasscutter (Thryonomys swinderianus rodentia: Hystricomorpha) in Ghana. *Open J. Animal Sci.* 3, 145–153. doi:10.4236/OJAS.2013.33022

Adenyo, C., Kayang, B. B., Owusu, E. H., Inoue, E., and Inoue-Murayama, M. (2017). Genetic diversity of grasscutter (Thryonomys swinderianus, Rodentia, Hystricomorpha) in Ghana based on microsatellite markers. *West Afr. J. Appl. Ecol.* 25, 1–15. doi:10.4314/wajae.v25i2

Adeola, A. C., Ommeh, S. C., Murphy, R. W., Wu, S. F., Peng, M. S., and Zhang, Y. P. (2015). Mitochondrial DNA variation of Nigerian domestic helmeted Guinea fowl. *Anim. Genet.* 46, 576–579. doi:10.1111/age.12324

Adeola, A. C., Sola-Ojo, F. E., Opeyemi, Y. A., Oguntunji, A. O., Nneji, L. M., Ewuola, M. K., et al. (2022). Genetic diversity and population structure of muscovy duck (*Cairina moschata*) from Nigeria. *PeerJ* 10, e13318. doi:10.7717/peerj.13236

Adu, E. K., Alhassan, W. S., and Nelson, F. S. (1999). Smallholder farming of the greater cane rat, Thryonomys swinderianus, temminck, in southern Ghana: A baseline survey of management practices. *Trop. Anim. Health Prod.* 31, 223–232. doi:10.1023/A:1005267110830

Adu, E. K., Aning, K. G., Wallace, P. A., and Ocloo, T. O. (2000). Reproduction and mortality in a colony of captive greater cane rats, Thryonomys swinderianus, Temminck. *Trop. Anim. Health Prod.* 32, 11–17. doi:10.1023/A:1005284817764

Adu, E. K., Asafu-Adjaye, A., Hagan, B. A., and Nyameasem, J. K. (2017). The grasscutter: An untapped resource of Africa's grasslands. *Livest. Res. Rural. Dev.* 29. Available at: http://www.lrrd.org/lrrd29/3/jnya29047.html (Accessed August 10, 2021).

Agaviezor, B. O., Adefenwa, M. A., Peters, S. O., Yakubu, A., Adebambo, O. A., Ozoje, M. O., et al. (2012). Genetic diversity analysis of the mitochondrial D-loop of Nigerian indigenous sheep. *Anim. Genet. Resour. génétiques Anim. genéticos Anim.* 50, 13–20. doi:10.1017/s2078633612000070

Akeberegn, D., Getabalew, M., Getahun, D., Alemneh, T., and Zewdie, D. (2019). Importance of hybrid vigor or heterosis for animal breeding. *Biochem. Biotechnol. Res.* 7, 1–4. Available at: https://www.researchgate.net/publication/334416265 (Accessed August 19, 2021).

Akinwole, M. T., and Babarinde, I. A. (2019). Assessing tissue lysis with sodium dodecyl sulphate for DNA extraction from frozen animal tissue. *J. Forensic Res.* 10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389

Aluko, F. A., Salako, A. E., Ngere, L. O., and Eniolorunda, O. O. (2015). Grasscutter: A review of the habitat, feeds and feeding, behaviour and economic importance. *Am. J. Res. Commun.* 3, 96–107. Available at: www.usa-journals.com (Accessed August 11, 2021).

Andem, J. A. (2012). Assessment of grasscutters ' (Thryonomys Swinderianus) sellers and hunters conservation knowledge, rate of hunting and methods of hunting in Oyo State. *Niger. Scholars Res. Libr.* 1, 86–92.

Annor, S. Y., Kagya-Agyemang, J. K., Abbam, J. E. Y., Oppong, S. K., and Agoe, I. M. (2008). Growth performance of grasscutter (Thryonomys swinderianus) eating leaf and stem fractions of Guinea grass (*Panicum maximum*). *Livest. Res. Rural. Dev.* 20. Available at: http://www.lrrd.org/lrrd20/8/anno20125.htm (Accessed August 10, 2021).

Babarinde, I. A., and Saitou, N. (2013). Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. *Genome Biol. Evol.* 5, 2330–2343. doi:10.1093/gbe/evt177

Babarinde, I. A., and Saitou, N. (2020). The dynamics, causes, and impacts of mammalian evolutionary rates revealed by the analyses of capybara draft genome sequences. *Genome Biol. Evol.* 12, 1444–1458. doi:10.1093/gbe/evaa157

Bandelt, H. J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48. doi:10.1093/OXFORDJOURNALS.MOLBEV.A026036

Baptist, R., and Mensah, G. A. (1986a). Benin and West Africa: The cane rat - farm animal of the future. *Anim. World* 60, 2–6.

Baptist, R., and Mensah, G. A. (1986b). The cane rat. Farm animal of the future. *World Rev. Anim. Prod.* 60, 2–6.

Barido-Sottani, J., Bošková, V., Plessis, L., Du, Kühnert, D., Magnus, C., Mitov, V., et al. (2018). Taming the BEAST—a community teaching material resource for BEAST 2. *Syst. Biol.* 67, 170–174. doi:10.1093/SYSBIO/SYX060

Bate, D. M. A. (1947). III—an extinct reed-rat (Thryonomys arkelli) from the Sudan. *Ann. Mag. Nat. Hist.* 14, 65–71. doi:10.1080/00222934708654610

Birchler, J. A., Yao, H., and Chudalayandi, S. (2006). Unraveling the genetic basis of hybrid vigor. *Proc. Natl. Acad. Sci.* 103, 12957–12958. doi:10.1073/PNAS.0605627103

Bouckaert, R. R. (2022). An efficient coalescent epoch model for bayesian phylogenetic inference. *Syst. Biol.* 71, 1549–1560. doi:10.1093/SYSBIO/SYAC015

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). Beast 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15, e1006650. doi:10.1371/JOURNAL.PCBI.1006650

Child, M. F. (2016). Thryonomys swinderianus. IUCN red list threat. *Species* 2016, e.T21847A115163896. doi:10.2305/IUCN.UK.2016-3.RLTS.T21847A22278009.en8235

Coker, O. M., Omonona, A. O., Fagbohun, O. A., Pylant, C., and Austin, J. D. (2017). Genetic structure of wild and domesticated grasscutters (Thryonomys swinderianus) from South-western Nigeria. *Afr. Zool.* 52, 155–162. doi:10.1080/15627020.2017.1379358

D'Elía, G., Fabre, P. H., and Lessa, E. P. (2019). Rodent systematics in an age of discovery: Recent advances and prospects. *J. Mammal.* 100, 852–871. doi:10.1093/JMAMMAL/GYY179

Douglas, J., Zhang, R., and Bouckaert, R. (2021). Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLoS Comput. Biol.* 17, e1008322. doi:10.1371/JOURNAL.PCBI.1008322

Durowaye, A. K., Salako, A. E., Osaiyuwu, O. H., and Fijabi, O. E. (2021). Relationship among liveweight and body dimensions of the greater cane rat (thrynomys swinderianus). *Asian J. Res. Anim. Vet. Sci.* 8, 1–9.

Essuman, E. K., and Duah, K. K. (2020). Poisonous substances used to capture and kill the greater cane rat (Thryonomys swinderianus). *Vet. Med. Sci.* 6, 617–622. doi:10.1002/VMS3.259

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

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491. doi:10.1093/GENETICS/131.2.479

Fabre, P. H., Hautier, L., Dimitrov, D., and P Douzery, E. J. (2012). A glimpse on the pattern of rodent diversification: A phylogenetic approach. *BMC Evol. Biol.* 12, 88. doi:10.1186/1471-2148-12-88

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., et al. (2019). Tracking five millennia of horse management with extensive ancient genome time series. *Cell* 177, 1419–1435. e31. doi:10.1016/J.CELL.2019.03.049/ATTACHMENT/3BDBAF9F-617E-4E35-AA60-357BBEC42C1F/MMC7.XLSX

Fayenuwo, J. . O., and Akande, M. (2002). The economic importance and control of cane-rat (Thryonomys swinderianus Temminck). *Proc. Vertebr. Pest Conf.* 20, 86–90. doi:10.5070/v420110171

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39, 783–791. doi:10.1111/J.1558-5646.1985.TB00420.X

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi:10.1007/BF01734359

Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147 (2), 915–925. doi:10.1093/genetics/147.2.915

Gaubert, P., Njiokou, F., Olayemi, A., Pagani, P., Dufour, S., Danquah, E., et al. (2015). Bushmeat genetics: Setting up a reference framework for the DNA typing of African forest bushmeat. *Mol. Ecol. Resour.* 15, 633–651. doi:10.1111/1755-0998.12334

Hajduk, G. K., Cockburn, A., Margraf, N., Osmond, H. L., Walling, C. A., and Kruuk, L. E. B. (2018). Inbreeding, inbreeding depression, and infidelity in a cooperatively breeding bird. *Evolution* 72, 1500–1514. doi:10.1111/EVO.13496

Harpending, H. C. (1994). Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum. Biol.* 66, 591–600.

Harrisson, K. A., Magrath, M. J. L., Yen, J. D. L., Pavlova, A., Murray, N., Quin, B., et al. (2019). Lifetime fitness costs of inbreeding and being inbred in a critically endangered bird. *Curr. Biol.* 29, 2711–2717. e4. doi:10.1016/j.cub.2019.06.064

Hasegawa, M., Kishino, H., and Yano, T. (2005). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *undefined* 22, 160–174. doi:10.1007/BF02101694

Heled, J., and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27, 570–580. doi:10.1093/MOLBEV/MSP274

Hoffmann, M. (2008). *Thryonomys swinderianus [greater cane rat}*. United Kingdom: IUCN Red List Threat, 1–6.

Huchon, D., and Douzery, E. J. P. (2001). From the old world to the new world: A molecular chronicle of the phylogeny and biogeography of hystricognath rodents. *Mol. Phylogenet. Evol.* 20, 238–251. doi:10.1006/mpev.2001.0961

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi:10.1093/BIOINFORMATICS/BTN129

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.–15

Jori, F., Mensah, G. A., and Adjanohoun, E. (1995). Grasscutter production: An example of rational exploitation of wildlife. *Biodivers. Conserv.* 4, 257–265. doi:10.1007/BF00055972

Kalu, C., and Aiyeloja, A. A. (2002). Bushmeat marketing in Nigeria: A case study of Benin city and its environs. *ASSET - Ser. A Agric. Environ.* 2, 33–38.

Kardos, M., Åkesson, M., Fountain, T., Flagstad, Ø., Liberg, O., Olason, P., et al. (2018). Genomic consequences of intensive inbreeding in an isolated Wolf population. *Nat. Ecol. Evol.* 2, 124–131. doi:10.1038/s41559-017-0375-4

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., and Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evol. Appl.* 9, 1205–1218. doi:10.1111/EVA.12414

Kilwanila, S. I., Msalya, G. M., Lyimo, C. M., and Rija, A. A. (2021). Geographic biases in cane rat (thryonomyds) research may impede broader wildlife utilization and conservation in africa: A systematic review. *Sci. Afr.* 12, e00785. doi:10.1016/J.SCIAF.2021.E00785

Kraatz, B. P., Bibi, F., Hill, A., and Beech, M. (2013). A new fossil thryonomyid from the Late Miocene of the United Arab Emirates and the origin of African cane rats. *Naturwissenschaften* 100, 437–449. doi:10.1007/s00114-013-1043-4

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054

Lanfear, R., Calcott, B., Ho, S. Y. W., and Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29, 1695–1701. doi:10.1093/MOLBEV/MSS020

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi:10.1093/BIOINFORMATICS/BTM404

Lasagna, E., Ceccobelli, S., Cardinali, I., Perini, F., Bhadra, U., Thangaraj, K., et al. (2020). Mitochondrial diversity of Yoruba and fulani chickens: A biodiversity reservoir in Nigeria. *Poult. Sci.* 99, 2852–2860. doi:10.1016/j.psj.2019.12.066

Librado, P., and Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi:10.1093/BIOINFORMATICS/BTP187

López-Antoñanzas, R., Sen, S., and Mein, P. (2004). Systematics and phylogeny of the cane rats (Rodentia: Thryonomyidae). *Zool. J. Linn. Soc.* 142, 423–444. doi:10.1111/j.1096-3642.2004.00136.x

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.* 46, 523–536. doi:10.1093/SYSBIO/46.3.523

Mau, B., and Newton, M. A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 6, 122–131. doi:10.1080/10618600.1997.10474731

Mauki, D. H., Adeola, A. C., Ng'ang'a, S. I., Tijjani, A., Akanbi, I. M., Sanke, O. J., et al. (2021). Genetic variation of Nigerian cattle inferred from maternal and paternal genetic markers. *PeerJ* 9, e10607–e10622. doi:10.7717/peerj.10607

Merwe, M. (2015). Discriminating between Thryonomys swinderianus and Thryonomys gregorianus. *Afr. Zool.* 42, 165–171. doi:10.1080/15627020.2007.11407393

Mouchaty, S. K., Catzeflis, F., Janke, A., and Arnason, U. (2001). Molecular evidence of an African Phiomorpha-South American Caviomorpha clade and support for Hystricognathi based on the complete mitochondrial genome of the cane rat (Thryonomys swinderianus). *Mol. Phylogenet. Evol.* 18, 127–135. doi:10.1006/mpev.2000.0870

Mustapha, O. A., Teriba, E. E., Ezekiel, O. S., Olude, A. M., Akinloye, A. K., and Olopade, J. O. (2020). A study of scientific publications on the greater cane rat (Thryonomys swinderianus, Temminck 1827). *Anim. Model. Exp. Med.* 3, 40–46. doi:10.1002/AME2.12103

Opara, M. N. (2010). The grasscutter I: A livestock of tomorrow. *Res. J. For.* 4, 119–135. doi:10.3923/RJF.2010.119.135

Owusu, B. A., Adu, E. K., Awotwi, E. K., and Awumbila, B. (2010). Embryonic resorption, litter size and sex ratio in the grasscutter, Thryonomys swinderianus. *Anim. Reprod. Sci.* 118, 366–371. doi:10.1016/j.anireprosci.2009.08.013

Patterson, B. D., and Upham, N. S. (2014). A newly recognized family from the horn of africa, the heterocephalidae (rodentia: Ctenohystrica). *Zool. J. Linn. Soc.* 172, 942–963. doi:10.1111/zoj.12201

Phillips, M. J. (2015). Four mammal fossil calibrations: Balancing competing palaeontological and molecular considerations. *Palaeontol. Electron.* 18, 1–16. doi:10.26879/490

Poux, C., Chevret, P., Huchon, D., De Jong, W. W., and Douzery, E. J. P. (2006). Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. *Syst. Biol.* 55, 228–244. doi:10.1080/10635150500481390

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi:10.1093/GENETICS/155.2.945

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi:10.1093/SYSBIO/SYY032

Rannala, B., and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43, 304–311. doi:10.1007/BF02338839

Raymond, M., and Rousset, F. (1995). An exact test for population differentiation. *Evol. (N. Y).* 49, 1280–1283. doi:10.1111/j.1558-5646.1995.tb04456.x

Rogers, A. R., and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9, 552–569. doi:10.1093/OXFORDJOURNALS.MOLBEV.A040727

Saitou, N., and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Sallam, H. M., and Seiffert, E. R. (2016). New phiomorph rodents from the latest Eocene of Egypt, and the impact of Bayesian "clock"-based phylogenetic methods on estimates of basal hystricognath relationships and biochronology. *PeerJ* 4, e1717. doi:10.7717/peerj.1717

Sallam, H. M., and Seiffert, E. R. (2020). Revision of oligocene "paraphiomys" and an origin for crown thryonomyoidea (rodentia: Hystricognathi: Phiomorpha) near the oligocene-miocene boundary in afAfrica*Zool. J. Linn. Soc.* 190, 352–371. doi:10.1093/zoolinnean/zlz148

Sallam, H. M., Seiffert, E. R., Steiper, M. E., and Simons, E. L. (2009). Fossil and molecular evidence constrain scenarios for the early evolutionary and biogeographic history of hystricognathous rodents. *Proc. Natl. Acad. Sci. U. S. A.* 106, 16722–16727. doi:10.1073/pnas.0908702106

Sambrook, J., and Russell, D. W. (2001). *Molecular cloning: A laboratory manual*. 3rd ed. New York: Cold Spring Harbor Laboratory Press.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* 1, 126. doi:10.1038/S41559-017-0126

Sheng, G., Hu, J., Tong, H., Llamas, B., Yuan, J., Hou, X., et al. (2020). Ancient DNA of northern China Hystricidae sub-fossils reveals the evolutionary history of old world porcupines in the Late Pleistocene. *BMC Evol. Biol.* 20, 88. doi:10.1186/S12862-020-01656-X

Suh, A., Smeds, L., and Ellegren, H. (2015). The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol.* 13, 1002224. doi:10.1371/JOURNAL.PBIO.1002224

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi:10.1093/genetics/123.3.585

Takezaki, N., Nei, M., and Tamura, K. (2010). POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with windows interface. *Mol. Biol. Evol.* 27, 747–752. doi:10.1093/molbev/msp312

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi:10.1093/OXFORDJOURNALS.MOLBEV.A040023

Teye, M., Fuseini, A., and Odoi, F. N. A. (2020). Consumer acceptance, Carcass and sensory characteristics of meats of farmed and wild cane rats (Thryonomys swinderianus). *Sci. Afr.* 8, e00461. doi:10.1016/j.sciaf.2020.e00461

Timberlake, W. E. (2013). "Heterosis," in *Brenner's encycl. Genet.* Second Ed. (Massachusetts: Academic Press), 451–453. doi:10.1016/B978-0-12-374984-0.00705-1

Tuomi, J. (1980). Mammalian reproductive strategies: A generalized relation of litter size to body size. *Oecologia* 45, 39–44. doi:10.1007/BF00346705

Upham, N. S., and Patterson, B. D. (2015). Evolution of the caviomorph rodents: A complete phylogeny and time tree of living genera project. *Biol. caviomorph rodents Divers. Evol.* 1, 63–120. Available at: https://www.researchgate.net/publication/282577627.

Van Der Merwe, M. (1999). Breeding season and breeding potential of the greater cane rat (Thryonomys swinderianus) in captivity in South Africa. *J. Zool.* 34, 69–73. doi:10.1080/02541858.1999.11448490

Visser, J. H., Bennett, N. C., and Jansen van Vuuren, B. (2019). Evolutionary and ecological patterns within the South African Bathyergidae: Implications for taxonomy. *Mol. Phylogenet. Evol.* 130, 181–197. doi:10.1016/J.YMPEV.2018.10.017

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, 1860–1869. doi:10.1038/ncomms2814

Wang, G. D., Zhai, W., Yang, H. C., Wang, L., Zhong, L., Liu, Y. H., et al. (2015). Out of southern east asia: The natural history of domestic dogs across the world. *Cell Res.* 26, 21–33. doi:10.1038/cr.2015.147

Wilson, D. E., and Reeder, D. A. M. (2005). *Mammal species of the world: A taxonomic and geographic reference*. Maryland, United States: Johns Hopkins University Press.

Woods, C. A., and Kilpatrick, C. W. (2005). "Infraorder hystricognathi," in *Mammal species of the world: A taxonomic and geographic referenc*. Editors D. E. Wilson and D. M. Reeder (Maryland, United States: Johns Hopkins University Press), 1545.

Yisau, M. A., Osunsina, I. O. O., and Onadeko, S. A. (2019). Wildlife population, structure and reproduction based on hunters' returns to a bush meat market in Abeokuta, Ogun state, Nigeria. *Adv. For. Sci.* 6, 541. doi:10.34062/afs.v6i1.7467

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* 35, 504–517. doi:10.1093/MOLBEV/MSX307

TYPE Brief Research Report
PUBLISHED 16 March 2023
DOI 10.3389/fevo.2023.1084502

Deleterious mutation load in the admixed mice population

Umayal Ramasamy[1,2], Abigail Elizur[2] and Sankar Subramanian[1,2]*

[1]School of Science, Engineering, and Technology, The University of the Sunshine Coast, Moreton Bay, QLD, Australia, [2]Centre for Bioinnovation, The University of the Sunshine Coast, Sippy Downs, QLD, Australia

Check for updates

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Diego Ortega-Del Vecchyo,
National Autonomous University of Mexico,
Mexico
Ana Yansi Morales Arce,
University of Bern,
Switzerland

*CORRESPONDENCE
Sankar Subramanian
✉ ssankara@usc.edu.au

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 30 October 2022
ACCEPTED 28 February 2023
PUBLISHED 16 March 2023

CITATION
Ramasamy U, Elizur A and Subramanian S (2023) Deleterious mutation load in the admixed mice population. *Front. Ecol. Evol.* 11:1084502. doi: 10.3389/fevo.2023.1084502

Deleterious mutation loads are known to correlate negatively with effective population size (N_e). Due to this reason, previous studies observed a higher proportion of harmful mutations in small populations than that in large populations. However, the mutational load in an admixed population that derived from introgression between individuals from two populations with vastly different N_e is not known. We investigated this using the whole genome data from two subspecies of the mouse (*Mus musculus castaneus* and *Mus musculus musculus*) with significantly different N_e. We used the ratio of diversities at nonsynonymous and synonymous sites (dN/dS) to measure the harmful mutation load. Our results showed that this ratio observed for the admixed population was intermediate between those of the parental populations. The dN/dS ratio of the hybrid population was significantly higher than that of *M. m. castaneus* but lower than that of *M. m. musculus*. Our analysis revealed a significant positive correlation between the proportion of *M. m. musculus* ancestry in admixed individuals and their dN/dS ratio. This suggests that the admixed individuals with high proportions of *M. m. musculus* ancestry have large dN/dS ratios. We also used the proportion of deleterious nonsynonymous SNVs as a proxy for deleterious mutation load, which also produced similar results. The observed results were in concordance with those expected by theory. We also show a shift in the distribution of fitness effects of nonsynonymous SNVs in the admixed genomes compared to the parental populations. These findings suggest that the deleterious mutation load of the admixed population is determined by the proportion of the ancestries of the subspecies. Therefore, it is important to consider the status and the level of genetic admixture of the populations whilst estimating the mutation loads.

KEYWORDS

mutation load, deleterious SNVs, mouse, population bottleneck, genetic admixture, introgression

Introduction

Population genetic theories predict a negative relationship between effective population size (N_e) and the fraction of deleterious mutations (Kimura and Ohta, 1978; Kimura, 1983). According to this, populations with small N_e accumulate more deleterious mutations than those with large N_e. This is because the effect of genetic drift is high amongst small populations, and hence the purifying selection is inefficient in purging deleterious mutations. Initially, accumulation of deleterious mutations was assumed to occur only in the asexual population, which was called as "Muller's ratchet" (Muller, 1964; Felsenstein, 1974). Later, Lynch et al. (1995) showed the possibility of accumulation of harmful mutations in sexual populations, particularly when the population is small. The proportion of deleterious mutations has been estimated in a number of species, including humans, chimpanzees, mice, cows, chickens, fruit flies, and plants (Eyre-Walker and Keightley, 1999; Schultz et al., 1999; Li and Saunders, 2005; Rubin et al., 2010; Bourneuf et al., 2017; Sohail, et al., 2017). These studies showed that amongst mammals, humans have the highest proportion of deleterious mutations. This is because the ancestral effective population size of humans is known to be the smallest of the mammals examined so far (Keightley and Eyre-Walker, 2000). Comparisons involving different populations within a species also showed similar patterns. For example, previous studies reported that European Americans have a significantly higher proportion of deleterious mutations than African Americans (Lohmueller et al., 2008; Fu et al., 2014; Subramanian, 2016). This observation could be explained based on the population bottleneck that occurred in Europeans when they migrated out of Africa. This was further confirmed by another large-scale study that showed a positive correlation between deleterious mutation load and the distance from Africa (Henn et al., 2016). Human populations experienced serial population bottlenecks as they migrated out of Africa, and populations that are furthest from Africa might have experienced more bottlenecks than those proximal to Africa. Therefore, the latter harbour more deleterious mutations than the former (Henn et al., 2016).

The effects of bottlenecks were also observed in populations that migrated much more recently. For instance, French Canadians have more deleterious mutations than those who live in France or Non-French Canadians (Peischl et al., 2018). Similarly, people living on Islands such as Greenland were also found to have more deleterious mutations than their mainland counterparts (Pedersen et al., 2017). In both cases, severe bottlenecks occurred due to the limited number of individuals who formed the founders of these populations. Hence the reduction in the population size caused the accumulation of higher proportions of deleterious mutations in French Canadians and Greenlanders than in their mainland counterparts.

During domestication, only a small subset of wild animals was selected, and hence the effective population size of domesticated animals is significantly reduced compared to their wild counterparts (Moyers et al., 2018). Due to this reason, the proportion of deleterious mutations is expected to be higher in domesticated animals than in their wild progenitors. A number of studies provided empirical evidence for this prediction. For example, an earlier study showed that domesticated dog breeds have a much higher deleterious mutation load than wild wolves (Marsden et al., 2016). Similar results were reported by comparing the exomes of domesticated and wild yaks (Xie et al., 2018). Using 432 animals belonging to 54 worldwide cattle breeds, a previous study found a negative correlation between heterozygosity and deleterious mutational load, which suggests that breeds that have low genetic variation have higher deleterious mutational load than those with high genetic variation (Subramanian, 2021). Since heterozygosity is the product of N_e and mutation rate (μ), and the mutation rate is largely similar across breeds, the above relationship implies a negative correlation between N_e and the mutation load.

Apart from domestication, the difference in the natural population sizes also cause a variation in the mutational loads. For instance, the island population of kakapo (a parrot species endemic to New Zealand) was found to carry a much higher proportion of deleterious variants than the mainland population (Foster et al., 2021). Similarly, the deleterious mutational load of mammoths that lived on Wrangel island was much higher than that of those who lived on the mainland (Rogers and Slatkin, 2017). Since the population size of island species are smaller than that of mainland ones, the accumulation of deleterious mutations was the result of high genetic drift modulating the mutations of the former. Similarly, killifish populations in drylands harboured more deleterious variants than those from the wetlands (Willemsen et al., 2020). This is because the former is known to have a smaller population size than the latter. Furthermore, a study on Alpine ibex found evidence for the accumulation of mildly deleterious SNVs but the purging of highly deleterious SNVs during population bottleneck (Grossen et al., 2020).

Almost all of the previous studies used the ratio of nonsynonymous-to-synonymous diversities (dN/dS) as the proxy for the deleterious mutation loads. Their findings clearly demonstrated low and high mutation loads in populations with large and small N_e. However, genetic admixture occurs quite often between populations with distinctly different N_e. For instance, a previous study showed the fitness of Neanderthals was significantly less than that of humans because the former has smaller N_e than the latter (Harris and Nielsen, 2016). Furthermore, this study revealed that non-African populations that admixed with Neanderthals (non-Africans) have a 0.5% reduction in fitness (Harris and Nielsen, 2016). However, it is important to understand how the mutational load of the admixed progenies derived from two populations (or subspecies) with widely different N_e (and, in turn, mutation loads) is shaped by evolution. To investigate this, we obtained the whole genome data of *Mus musculus castaneus* and *Mus musculus musculus* and their hybrid or admixed populations. We specifically selected these because the subspecies *M. m. musculus* and *M. m. castaneus* have population sizes of 60,000–120,000 and 200,000–400,000, respectively (Salcedo et al., 2007; Geraldes et al., 2008), and previous studies have found a much higher mutational load in the former compared to that observed in the latter (Phifer-Rixey et al., 2012; Subramanian, 2018). We estimated the dN/dS ratios and the proportion of deleterious nonsynonymous SNVs (nSNVs) for the admixed mouse population and compared them with those of their parental populations. We also examined the correlation between the level of admixture and deleterious mutational loads.

Materials and methods

Genome data

Whole genome data for 11 *Mus musculus castaneus*, 11 *Mus musculus musculus*, 18 hybrids (*M. m. castaneus*' *M. m. musculus*) and one *Mus spretus* were obtained from a previous study (Fujiwara et al., 2022). We only used the single nucleotide variations (excluding INDELS), and positions with more than two alleles were excluded. To determine derived alleles, the genomic positions of the outgroup *Mus spretus* were used to orient the mutations. We only included the positions where the outgroup was homozygous, and at least one *Mus musculus* genome had an allele that was different from the outgroup. We excluded the mitochondrial genome and sex chromosomes and used only autosomes. We used the genome annotations and extracted the protein-coding genes, and using the software PAML (Yang, 2007) we determined the number of nonsynonymous and synonymous sites. To identify nonsynonymous (nSNVs) and synonymous SNVs (sSNVs), the program *SNPEffect* was used (Cingolani et al., 2012).

Data analysis

The software *plink* was used to cluster the genomes based on their genetic relationship (Purcell et al., 2007), and the multidimensional scaling method was employed for this purpose. The VCF file was first converted into a binary bed format. The SNVs with <5% MAF (Minor Allele Frequency) were excluded, and genomes that contained >10% gap or unknown genotypes were excluded. Finally, the first two principal components that explain the maximum variation were used to plot the values and observe the population clusters. To determine the proportion of ancestries in each admixed genome of the hybrids the likelihood-based method *admixture* was used (Alexander et al., 2009). Using the cross-validation analysis, we determined the K value that had the lowest error and used that value (2) to denote the number of ancestral populations.

Using the number of nSNVs and sSNVs and their respective sites, we estimated the diversities at these sites and used them to compute their ratio (ω_{obs}) as:

$$\omega_{obs} = \frac{dN}{dS} \tag{1}$$

To determine the deleteriousness and to identify the harmful mutations, we used the *PhyloP* conservation scores (Hubisz et al., 2011). We obtained the basewise *PhyloP* scores based on 59 vertebrate genomes.[1] The score was available for each position of the mouse chromosomes. The *PhyloP* scores were then mapped to the genomic positions, and any SNV present in a genomic position with a PhyloP score > 2.0 was designated as deleterious in nature. Based on this, the nSNVs with a PhyloP score > 2.0 were considered deleterious nSNVs. The number of deleterious nSNVs (DN_{obs}) was divided by the number of all nSNVs (N_{obs}) to obtain the proportion of observed deleterious nSNVs (δ_{obs}) as given below:

$$\delta_{obs} = \frac{DN_{obs}}{N_{obs}} \tag{2}$$

To determine the distribution of fitness effects of the SNVs, the software DFE was used (Eyre-Walker et al., 2006). The site frequency spectrums (SFS) of *M. m. castaneus, M. m. musculus,* and the hybrid mice populations were obtained. To model the distribution, first the *lookuptable* program of this software was used. The output of this program was used along with the SFS of each population to obtain the fitness effects of the SNVs.

Theoretical prediction

To predict the genomic heterozygosity or diversity for admixed genomes (H_{adm}) by using the allele frequencies of their ancestral populations we used the equation developed by Boca et al. (2020), as given below:

$$d_{adm} = 1 - \sum_{j=1}^{N} \bar{p}_j^2 = 1 - \sum_{j=1}^{N} \left(\sum_{i=1}^{K} \gamma_i p_{ij} \right)^2 \tag{3}$$

where p_j is the frequency of allele *j*, *N* is the number of alleles, γ_i is the proportion of admixture of ancestor *i,* and *K* is the number of ancestral populations contributed to the genetic makeup of the admixed populations. d_{adm} calculated for nonsynonymous (dN_{admn}) and synonymous sites (dS_{adms}) and their ratio of the predicted diversities (ω_{pre}) was computed as:

$$\omega_{pre} = \frac{dN_{adm}}{dS_{adm}} \tag{4}$$

Similarly, to predict the proportion of deleterious nonsynonymous SNVs (δ_{pre}) based on the nonsynonymous (N_i) and deleterious nonsynonymous SNVs (DN_i) of the ancestral populations, we used the following formula:

$$\delta_{pre} = \frac{DN_{pre}}{N_{pre}} \tag{5}$$

where

$$N_{pre} = \sum_{i=1}^{K} \gamma_i N_i \tag{6}$$

$$DN_{pre} = \sum_{i=1}^{K} \gamma_i DN_i \tag{7}$$

Statistical analyses

The average dN/dS ratios and the mean proportion of deleterious nSNVs, along with the standard errors, were estimated for each pure and admixed genome. The significance between the mean estimates was determined using the Z test. A regression and correlation analysis was performed to study the relationship between the proportion of genetic admixture and dN/dS ratios or the proportion of deleterious nSNVs. The nonparametric Spearman's rank correlation was used to determine the strength of the correlation.[2] Furthermore, using the parametrical Pearson correlation also produced similar strength of the relationship. The statistical significance of the Pearson correlation was determined by converting the correlation coefficient *r* to the normal deviation *Z*, and this was accomplished using the online software *r* to

[1] http://hgdownload.Soe.ucsc.edu/goldenPath/mm10/phyloP60way/

[2] https://www.wessa.net/rwasp_spearman.wasp

P.[3] In-house Perl scripts were used to map the *PhyloP* scores, calculate the point estimates and standard errors, and perform statistical tests.

Results

Population structure and admixture

To understand the genetic relationship between parental and admixed populations, the Principal Component Analysis (PCA) was performed. The results showed distinct clusters for parental populations (Figure 1A). The genomes of *M. m. musculus* are on the left and those of *M. m. castaneus* on the right ends of the plot. In contrast, the admixed individuals spread in between their parental populations. The hybrid genomes close to *M. m. musculus* suggest a high proportion of musculus ancestry and a low proportion of *castaneus* ancestry in their genome. Those close to *M. m. castaneus* suggest the opposite. To estimate the actual proportion of these ancestries, we used the maximum likelihood-based software *Admixture*. Figure 1B shows the unadmixed parental populations in single colours, which confirms the purity of these genomes. The admixed individuals are shown in two colours suggesting the levels of admixture, and the size of each colour on the columns indicates the proportions of the corresponding ancestry. The admixed mice genomes have 10%–95% of *M. m. musculus* ancestry.

The ratio of nonsynonymous and synonymous diversities

To estimate the deleterious mutation load in each genome, we first calculated the dN/dS ratio for each genome. The average estimate and standard error were computed for *M. m. castaneus, M. m. musculus*, and the admixed populations. Figure 2A shows the mean estimates were very distinct for the populations, and the differences between them were highly significant ($p < 0.00001$, using a Z test). The ratio of *M. m. musculus* was higher than that of *M. m. castaneus*, and importantly that observed for the admixed population was intermediate between the two. We then plotted the dN/dS ratios of each admixed individual against the proportion of *M. m. musculus* ancestry of these genomes (Figure 2B). Our results showed a highly significant ($rho = 0.89$, $p < 0.00001$, Spearman's rank correlation) positive correlation between the two variables. The dN/dS ratio was small for admixed individuals with less proportion of *M. m. musculus* ancestry, and it was high for those with a high fraction of the *musculus* ancestry.

The proportion of deleterious nSNVs

To further examine the harmful mutation loads, we used the proportion of deleterious nSNVs as the measure to quantify the load. We computed this using homozygous nSNVs, and heterozygous nSNVs and then combined these two to obtain the proportion for all deleterious nSNVs. As shown in Figure 3A, the mean estimate of the proportion of all deleterious nSNVs was the highest for *M. m. musculus*, lowest for *M. m. castaneus*, and intermediate for the admixed population. The differences between the mean estimates were also statistically significant (at least $p < 0.00001$). Similar results were observed for the proportion of deleterious homozygous and heterozygous nSNVs as well (Figures 3B,C). Although the differences between the mean estimates were statistically significant, the magnitude of the differences was high for the heterozygous and low for the homozygous SNVs.

To understand the relationship between the proportion of *M. m. musculus* admixture/ancestry and the deleterious mutation load, a correlation analysis was conducted. For this purpose, first, we plotted the proportion of *M. m. musculus* ancestry against the proportion of all deleterious nSNVs of each genome (Figure 4A). This produced a highly significant positive relationship ($rho = 0.91$, $p < 0.00001$), which suggests that the admixed individuals with high proportions of *M. m. musculus* ancestry also have high proportions of deleterious nSNVs and those with low proportions have less proportions of these SNVs. Similarly, highly significant positive relationships were also observed for the homozygous ($rho = 0.94$, $p < 0.00001$) and heterozygous SNVs ($rho = 0.81$, $p < 0.0001$; Figures 4B,C).

Expected and observed mutational loads in admixed populations

Previous results revealed the patterns of mutation load observed in admixed populations. In order to understand the patterns expected solely based on the allele frequencies of the ancestral parent populations, we used the Equation (3) developed by a previous study to estimate the expected heterozygosity (see Methods). Using this formula, we first calculated the predicted nonsynonymous (dN_{adm}) and synonymous (dS_{adm}) diversities and obtained the predicted ratio of these diversities (ω_{pre}). We then compared this ratio with that observed for the admixed genomes (ω_{obs}). The observed nonsynonymous (dN) and synonymous (dN) diversities were 11% smaller than those corresponding predicted diversities (dN_{adm} and dS_{adm}, respectively). However, the observed ratio of these diversities (ω_{obs}) was not statistically different from that expected (ω_{pre}; $p = 0.33$; Figure 5). We also developed an Equation (5) to estimate the predicted proportion of deleterious SNVs (δ_{pre}) in admixed genomes based on the level of contributions of their ancestral populations (see Methods). The proportion of deleterious SNVs observed for the admixed populations was compared (δ_{obs}) with the predicted proportion of these SNVs (Figure 5). Our results did not reveal any significant difference between the predicted and observed proportions ($p = 0.32$).

Distribution of fitness effects

To infer the fitness effects of nonsynonymous SNVs, we used the software DFE by providing the site frequencies of nonsynonymous and synonymous SNVS as input. This analysis produced the fraction of nSNVs belonging to different fitness effect categories. The results revealed that the *M. m. musculus* population has the highest fraction of mildly deleterious or nearly neutral ($Nes < 10$) nSNVs and lowest

[3] http://vassarstats.net/tabs_z.html

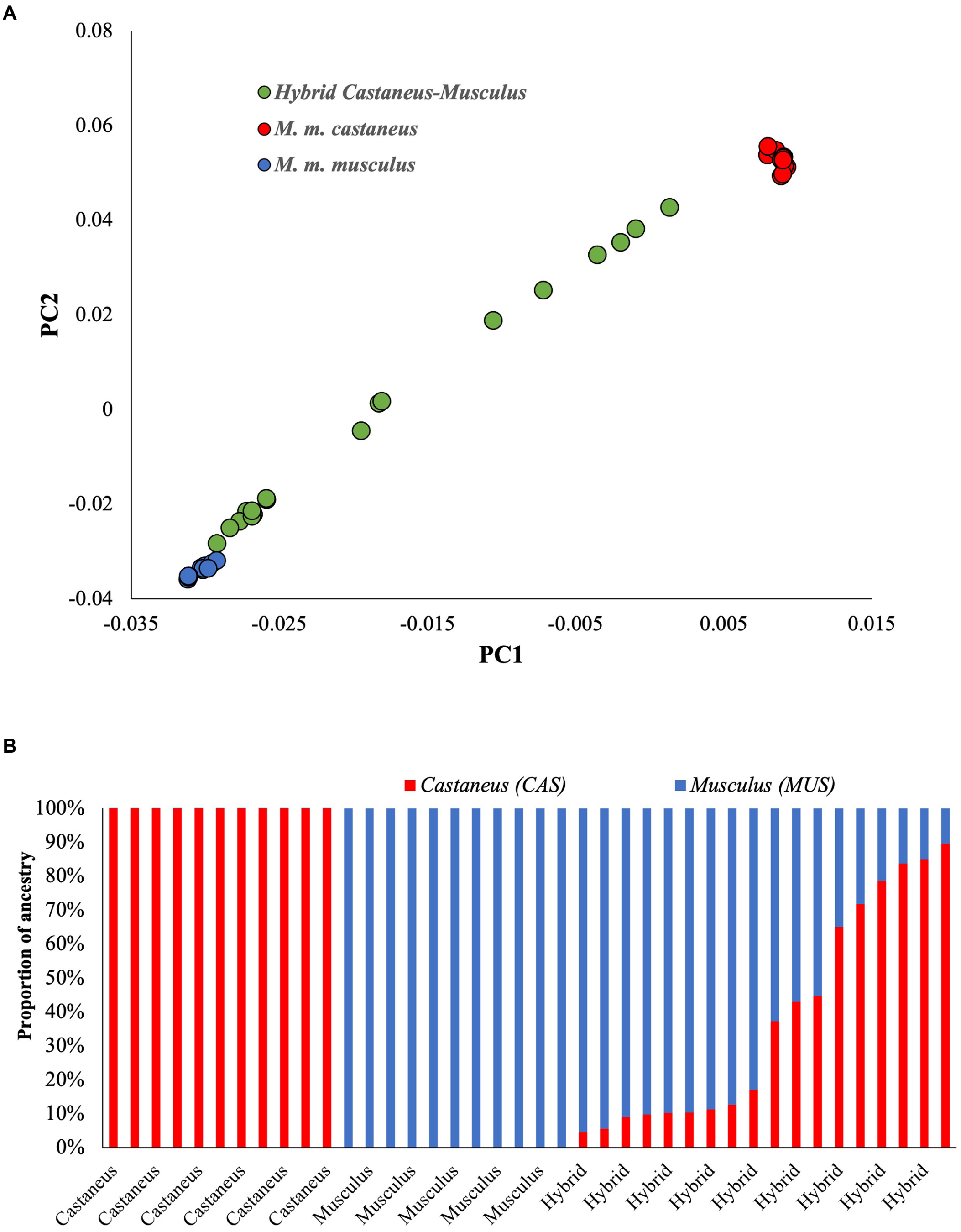


FIGURE 1
(A) Principal Component Analysis using pure and admixed mice populations. The pure populations (*Mus musculus castaneus* and *Mus musculus musculus*) are distinct and widely separated on the *x*-axis. In contrast, the admixed or hybrid populations spread between the parents. **(B)** The proportion of ancestries in each genome. Pure populations are shown in one colour, whereas the admixed individuals have two colours representing the proportions of different ancestries.

fractions of moderately (*Nes* 10–100) and highly deleterious nSNVs (*Nes* > 100; Figure 6). An opposite trend was observed for *M. m. castaneus* population. The admixed mice had a significantly higher fraction of mildly deleterious nSNVs ($p < 0.001$) than *M. m. castaneus* and significantly lower fraction of mildly deleterious nSNVs ($p < 0.001$) than *M. m. musculus* population. In contrast, the hybrid mice had a significantly lower fraction of moderately deleterious nSNVs ($p < 0.001$) than *M. m. castaneus* and a significantly higher fraction of moderately deleterious nSNVs ($p = 0.003$) than *M. m. musculus* population. Whilst the admixed populations had a significantly higher fraction of highly deleterious SNVs than *M. m. musculus* ($p < 0.001$) there was no significant difference in the fraction of highly deleterious SNVs between the admixed and *M. m. castaneus* populations ($p = 0.12$).

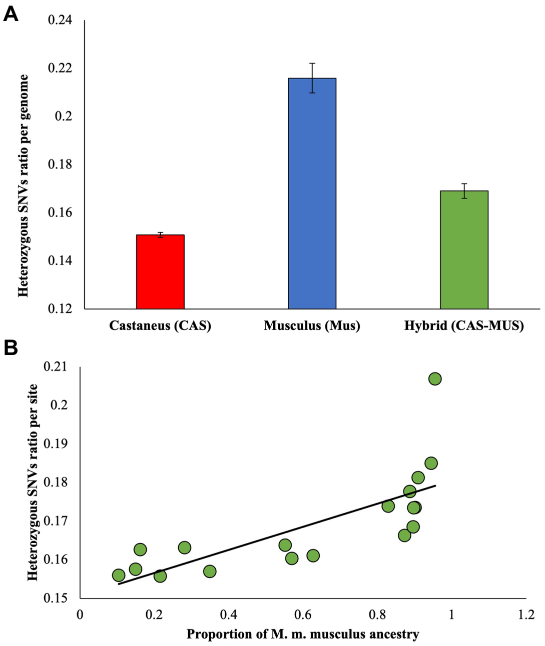


FIGURE 2
(A) The ratio of diversities at nonsynonymous and synonymous sites (ω_{obs}=*dN/dS*) was estimated for pure and admixed populations. The error bars show the standard error of the mean. The differences between *dN/dS* ratios of the hybrid and *M. m. musculus* (p<0.00001) and that between the hybrid and *M. m. castaneus* (p<0.00001) populations were significant. **(B)** Correlation between the proportion of *M. m. musculus* ancestry/admixture and the *dN/dS* ratio estimated for admixed individuals. The relationship was highly significant (*rho*=0.89, p<0.00001), and the best-fitting curve is shown.

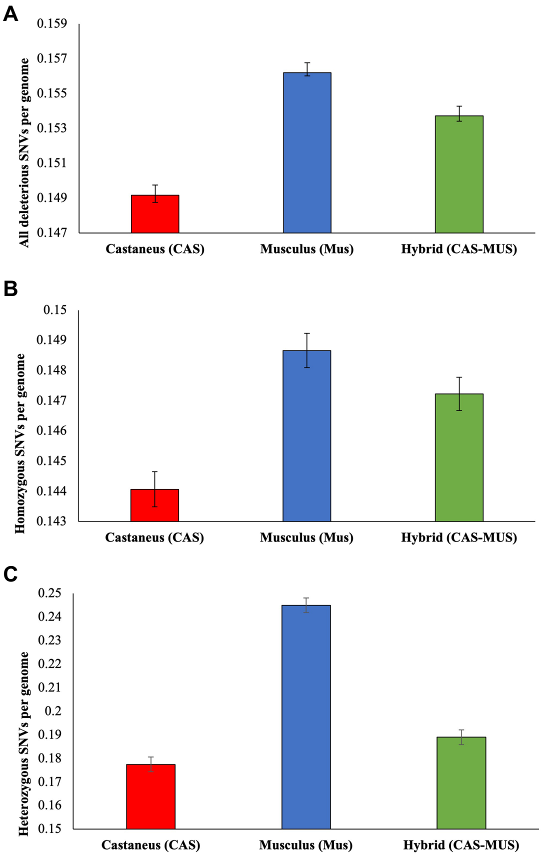


FIGURE 3
The proportion of deleterious nonsynonymous SNVs (nSNVs) (δ_{obs}) computed for *M. m. castaneus*, *M. m. musculus,* and admixed populations. **(A)** All deleterious nSNVs, **(B)** Homozygous deleterious nSNVs, and **(C)** Heterozygous deleterious nSNVs. The error bars show the standard error of the mean. The differences between the proportion of deleterious nSNVs of the hybrid and *M. m. musculus* were significant for all (p<0.00001), homozygous (p<0.0341), and heterozygous SNVs (p<0.00001). Similarly, the differences between the hybrid and *M. m. castaneus* were also significant for All (p<0.00001), homozygous (p<0.00001), and heterozygous SNVs (p<0.0001).

Discussion

Population genetic theories predict a higher proportion of deleterious mutations in a population with a small N_e compared to that with a large N_e (Crow and Kimura, 1970; Phifer-Rixey et al., 2012; Marsden et al., 2016; Subramanian, 2021). Many previous studies provided empirical evidence for this prediction using genome data from humans and other vertebrates (Lu et al., 2006; Mezmouk and Ross-Ibarra, 2014; Renaut and Rieseberg, 2015; Kono et al., 2016; Marsden et al., 2016; Pedersen et al., 2017; Ramu et al., 2017; Makino et al., 2018; Peischl et al., 2018; Xie et al., 2018; Bosse et al., 2019; Robinson et al., 2019; Dussex et al., 2021; Subramanian, 2021). Using the dN/dS ratio as the measure to quantify deleterious mutation load, these studies consistently showed that the populations with small N_e had a higher ratio than those with large N_e. However, the mutation load of the populations that derive from genetic admixture between two parental populations with distinctly different N_e is not known. In the present study, using the dN/dS ratio and the proportion of deleterious nSNVs as proxies to quantify deleterious mutation load in mouse populations that result from the admixture between the subspecies *M. m. castaneus* and *M. m. musculus*. We report two main findings. First, the mean dN/dS ratio and the proportion of deleterious nSNVs estimated for the admixed mouse population was intermediate between those of observed for their parental populations. Second, these measures positively correlated with the fraction of *M. m. musculus* ancestry (or negatively correlated with the fraction of *M. m. castaneus* ancestry).

The results of this study can be explained by the fact that the N_e of *M. m. castaneus* (200,000–400,000) is much higher than that of *M. m. musculus* (60,000–120,000; Salcedo et al., 2007; Geraldes et al., 2008). Since deleterious mutation load negatively correlates N_e (Kimura and Ohta, 1978; Kimura, 1983), this load is expected to be higher in *M. m. musculus* compared to that in *M. m. castaneus*. Our results exactly showed this, as the dN/dS ratio (Figure 2A) and the proportion of deleterious nSNVs (Figure 3) were found to be higher for the former than those for the latter. On the other hand, the deleterious mutation load for the population that derived from the admixture of these two subspecies is expected to be less than that of *M. m. musculus* as they have a fraction of *M. m. castaneus* ancestry, which will tend to reduce the overall load of the admixed individuals. Similarly, the deleterious mutation load of the admixed population is expected to be higher than that of *M. m. castaneus* as they have a fraction of *M. m. musculus* ancestry, which will tend to increase the overall load of the hybrids.

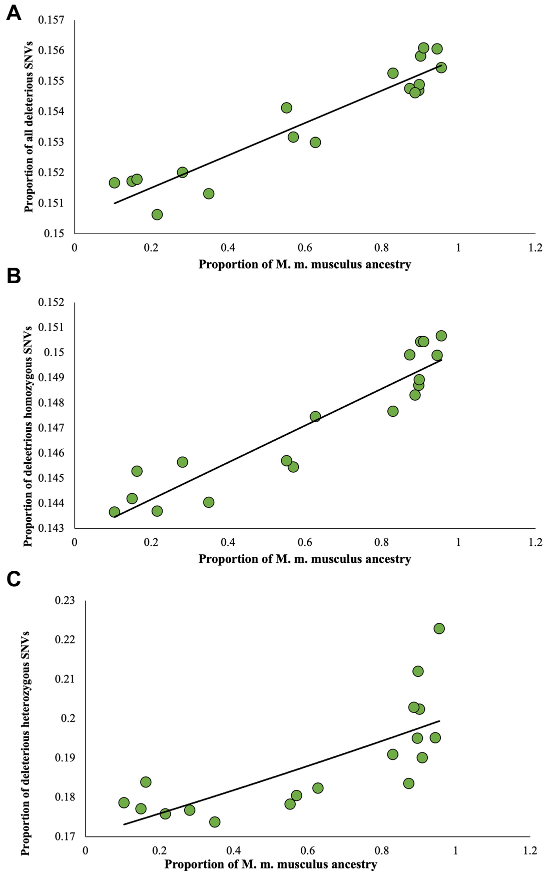


FIGURE 4
Relationship between the proportion of *M. m. musculus* ancestry/admixture and the proportion of deleterious nSNVs estimated for admixed individuals. **(A)** All deleterious nSNVs, **(B)** Homozygous deleterious nSNVs, and **(C)** Heterozygous deleterious nSNVs. The relations were highly significant for all (*rho*=0.91, *p*<0.00001), homozygous (*rho*=0.94, p<0.00001), and heterozygous SNVs (*rho*=0.81, p<0.0001). The best-fitting curves are shown.

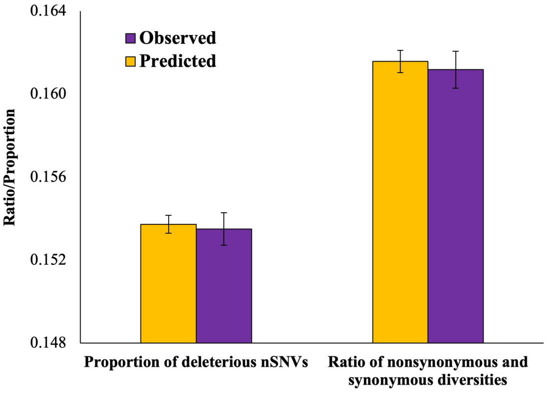


FIGURE 5
Column graph comparing the observed (ω_{obs}) and predicted (ω_{pre}) ratios of nonsynonymous and synonymous diversities and the proportion of deleterious SNVs observed (δ_{obs}) and predicted (δ_{pre}). The differences between the observed and predicted values of the ratios (*p*=0.33) and the proportions (*p*=0.32) were not statistically significant.

Due to this reason, the dN/dS ratio (Figure 2A) and the proportion of deleterious nSNVs (Figure 3) observed for admixed genomes were higher than those of *M. m. castaneus* and less than those of *M. m. musculus* population. The above prediction was further confirmed by the positive correlation between the proportion of *M. m. musculus* ancestry and the dN/dS ratio (Figure 2B) or the proportion of deleterious nSNVs (Figure 4). Note that if we used the *M. m. castaneus* ancestry we found a negative correlation with the same strength and same level of statistical significance.

Previous studies on dogs, cows, and yak used deleterious heterozygous and homozygous nSNVs counts to compare mutational loads between populations with different N_e (Marsden et al., 2016; Xie et al., 2018; Subramanian, 2021). However, populations with small N_e are expected to have a higher number of deleterious as well as neutral (or benign) homozygous nSNVs than those with large N_e. Therefore, the above studies observed a high number of deleterious and neutral homozygous SNVs in the genomes of the former and low in the latter. Since the deleterious and neutral homozygous SNVs show similar patterns, the use of SNV counts to measure the mutational load is not accurate. Due to this reason, we used the proportion of deleterious homozygous nSNVs, which was calculated by dividing the number of deleterious homozygous nSNVs by the number of all (benign + deleterious) homozygous nSNVs (Figures 3B, 4B) and this proportion removes the bias expected in using the nSNV counts. A similar measure was used for heterozygous deleterious nSNVs as well (Figures 3C, 4C).

Although we observed the ratio of nonsynonymous and synonymous diversities in the admixed genomes it is important to examine the loads expected in these based on the allele frequencies of their parental populations. Although the observed nonsynonymous and synonymous diversities were significantly smaller than those predicted, their ratios were not. This is because the rate of reduction in the observed estimates of both nonsynonymous and synonymous sites was almost the same (~11%). This result and the similarity between the predicted and observed proportion of deleterious nSNVs

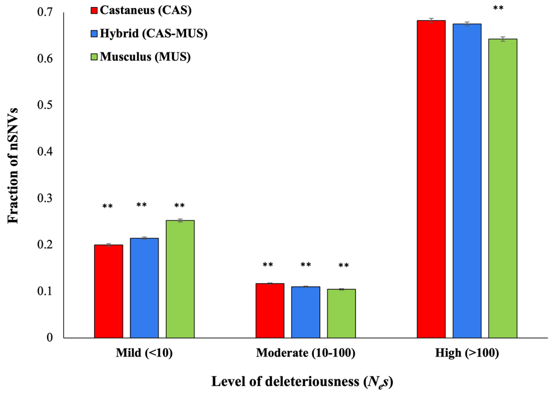


FIGURE 6
Distribution of fitness effects of nonsynonymous SNVs (nSNVs). The X-axis shows the level of deleteriousness of nSNVs in terms of the product of effective population size (N_e) and selection coefficient (*s*). The nSNVs were grouped into mildly (N_es<10), moderately (N_es 10–100) and highly (N_es>100) deleterious in nature. The double asterisks indicate statistical significance at the 1% level (*Z*-test, two-tailed). The actual *p*-values for each comparison are given in the main text (see Results).

suggest that the level of admixture more likely determines the mutational loads of the admixed progenies.

The distribution of fitness effects of nonsynonymous nSNVs revealed a higher proportion of mildly deleterious SNVs and a lower proportion of moderately and highly deleterious nSNVs in *M. m. musculus* than *M. m. castaneus*. This suggests that some of the highly deleterious nSNVs that were removed from *M. m. castaneus* population were present or segregating in the *M. m. musculus* population. This is because these nSNVs were deleterious for the former but became or behaved as mildly deleterious in the latter population. The potential reason is that the presence or removal of the nSNVs is determined by the product of effective population size (N_e) and selection coefficient (s). Therefore, for a large population, the N_es will be larger than that of a small population for the same vale of s. Hence, the fitness consequences of nSNVs in small populations will be relatively less deleterious than that in large populations. Therefore, some of the nSNVs that were highly deleterious for *M. m. castaneus* were mildly deleterious for *M. m. musculus*. The genetic Admixture between *M. m. castaneus* and *M. m. musculus* results in the loss of some of the moderately deleterious nSNVs whilst gaining or accumulating some of the mildly deleterious ones.

Although many previous studies estimated the mutation load in pure or unadmixed populations, how these loads are modulated by the proportion of ancestry in admixed genomes was not clearly demonstrated before. Hence, the analyses performed here make this study novel and unique. Due to this new approach, the results of this study have opened up new dimensions in understanding the diversity and mutational load of admixed populations in comparison to their ancestral/parental populations. Eventually, this approach will help to recognise the role of gene flow between populations, particularly in exchanging harmful mutations.

Conclusion

Using the whole genome data, this study revealed the patterns of deleterious mutation load in a mouse population, which derived from genetic admixture between two distinct populations belonging to the subspecies *M. m. castaneus* and *M. m. domesticus*. The results showed the dN/dS ratio and the proportion of deleterious nSNVs of admixed populations were intermediate compared to their parents. This could be observed only if the parental populations had significant variation in their mutational load, which is driven by their effective population sizes. Our study also revealed a significant positive correlation between the proportion of admixture and the dN/dS ratio or the proportion of deleterious nSNVs in admixed individuals. This suggests the role of admixture in shaping the deleterious mutation loads. Whilst this was observed in mice, it is highly likely that similar patterns are expected in other vertebrates as well. Therefore, it is important to consider the status and the level of genetic admixture of the populations whilst estimating the mutation loads.

Data availability statement

Publicly available datasets were analysed in this study. This data can be found at: https://www.ddbj.nig.ac.jp/bioproject/ accession: PRJDB11027.

Author contributions

SS conceived the idea, designed and supervised the research, and wrote the manuscript with inputs from other authors. UR performed research and data analysis, and contributed to writing the manuscript. AE supervised the research and contributed to writing and editing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The research was supported by the University of the Sunshine Coast DVCR grant awarded to SS.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

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

Boca, S. M., Huang, L., and Rosenberg, N. A. (2020). On the heterozygosity of an admixed population. *J. Math. Biol.* 81, 1217–1250. doi: 10.1007/s00285-020-01531-9

Bosse, M., Megens, H. J., Derks, M. F. L., Cara, Á. M. R., and Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. *Evol. Appl.* 12, 6–17. doi: 10.1111/eva.12691

Bourneuf, E., Otz, P., Pausch, H., Jagannathan, V., Michot, P., Grohs, C., et al. (2017). Rapid discovery of de novo deleterious mutations in cattle enhances the value of livestock as model species. *Sci. Rep.* 7:11466. doi: 10.1038/s41598-017-11523-3

Cingolani, P., Platts, A., Wang le, 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 (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Crow, J. F., and Kimura, M. (1970). *An introduction to population genetics theory. An introduction to population genetics theory.* New York: Harper & Row.

Dussex, N., van der Valk, T., Morales, H. E., Wheat, C. W., Díez-del-Molino, D., von Seth, J., et al. (2021). Population genomics of the critically endangered kākāpō. *Cell Genomics* 1:100002. doi: 10.1016/j.xgen.2021.100002

Eyre-Walker, A., and Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature* 397, 344–347. doi: 10.1038/16915

Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173, 891–900. doi: 10.1534/genetics.106.057570

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* 78, 737–756. doi: 10.1093/genetics/78.2.737

Foster, Y., Dutoit, L., Grosser, S., Dussex, N., Foster, B. J., Dodds, K. G., et al. (2021). Genomic signatures of inbreeding in a critically endangered parrot, the kakapo. *G3 (Bethesda)* 11:jkab307. doi: 10.1093/g3journal/jkab307

Fu, Q. M., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., et al. (2014). Genome sequence of a 45,000-year-old modern human from western siberia. *Nature* 514:445. doi: 10.1038/nature13810

Fujiwara, K., Kawai, Y., Takada, T., Shiroishi, T., Saitou, N., Suzuki, H., et al. (2022). Insights into mus musculus population structure across eurasia revealed by whole-genome analysis. *Genome Biol. Evol.* 14, 14:evac068. doi: 10.1093/gbe/evac068

Geraldes, A., Basset, P., Gibson, B., Smith, K. L., Harr, B., Yu, H. T., et al. (2008). Inferring the history of speciation in house mice from autosomal, x-linked, y-linked and mitochondrial genes. *Mol. Ecol.* 17, 5349–5363. doi: 10.1111/j.1365-294X.2008.04005.x

Grossen, C., Guillaume, F., Keller, L. F., and Croll, D. (2020). Purging of highly deleterious mutations through severe bottlenecks in alpine ibex. *Nat. Commun.* 11:1001. doi: 10.1038/s41467-020-14803-1

Harris, K., and Nielsen, R. (2016). The genetic cost of neanderthal introgression. *Genetics* 203, 881–891. doi: 10.1534/genetics.116.186890

Henn, B. M., Botigue, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., et al. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl. Acad. Sci. U. S. A.* 113, E440–E449. doi: 10.1073/pnas.1510805112

Hubisz, M. J., Pollard, K. S., and Siepel, A. (2011). Phast and rphast: phylogenetic analysis with space/time models. *Brief. Bioinform.* 12, 41–51. doi: 10.1093/bib/bbq072

Keightley, P. D., and Eyre-Walker, A. (2000). Deleterious mutations and the evolution of sex. *Science* 290, 331–333. doi: 10.1126/science.290.5490.331

Kimura, M. (1983). *The neutral theory of molecular evolution Cambridge*, UK: Cambridge University Press.

Kimura, M., and Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U. S. A.* 75, 2868–2872. doi: 10.1073/pnas.75.6.2868

Kono, T. J., Fu, F., Mohammadi, M., Hoffman, P. J., Liu, C., Stupar, R. M., et al. (2016). The role of deleterious substitutions in crop genomes. *Mol. Biol. Evol.* 33, 2307–2317. doi: 10.1093/molbev/msw102

Li, W. H., and Saunders, M. A. (2005). News and views: the chimpanzee and us. *Nature* 437, 50–51. doi: 10.1038/437050a

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451, 994–997. doi: 10.1038/nature06611

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S., and Wu, C. I. (2006). The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* 22, 126–131. doi: 10.1016/j.tig.2006.01.004

Lynch, M., Conery, J., and Burger, R. (1995). Mutational meltdowns in sexual populations. *Evolution* 49, 1067–1080. doi: 10.1111/j.1558-5646.1995.tb04434.x

Makino, T., Rubin, C. J., Carneiro, M., Axelsson, E., Andersson, L., and Webster, M. T. (2018). Elevated proportions of deleterious genetic variation in domestic animals and plants. *Genome Biol. Evol.* 10, 276–290. doi: 10.1093/gbe/evy004

Marsden, C. D., Ortega-Del Vecchyo, D., O'Brien, D. P., Taylor, J. F., Ramirez, O., Vilà, C., et al. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc. Natl. Acad. Sci. U. S. A.* 113, 152–157. doi: 10.1073/pnas.1512501113

Mezmouk, S., and Ross-Ibarra, J. (2014). The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)* 4, 163–171. doi: 10.1534/g3.113.008870

Moyers, B. T., Morrell, P. L., and McKay, J. K. (2018). Genetic costs of domestication and improvement. *J. Hered.* 109, 103–116. doi: 10.1093/jhered/esx069

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat. Res.* 106, 2–9. PMID: 14195748

Pedersen, C. T., Lohmueller, K. E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H. R., et al. (2017). The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: insights from the greenlandic inuit. *Genetics* 205, 787–801. doi: 10.1534/genetics.116.193821

Peischl, S., Dupanloup, I., Foucal, A., Jomphe, M., Bruat, V., Grenier, J. C., et al. (2018). Relaxed selection during a recent human expansion. *Genetics* 208, 763–777. doi: 10.1534/genetics.117.300551

Phifer-Rixey, M., Bonhomme, F., Boursot, P., Churchill, G. A., Pialek, J., Tucker, P. K., et al. (2012). Adaptive evolution and effective population size in wild house mice. *Mol. Biol. Evol.* 29, 2949–2955. doi: 10.1093/molbev/mss105

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

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J. V., et al. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* 49, 959–963. doi: 10.1038/ng.3845

Renaut, S., and Rieseberg, L. H. (2015). The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Mol. Biol. Evol.* 32, 2273–2283. doi: 10.1093/molbev/msv106

Robinson, J. A., Raikkonen, J., Vucetich, L. M., Vucetich, J. A., Peterson, R. O., Lohmueller, K. E., et al. (2019). Genomic signatures of extensive inbreeding in isle royale wolves, a population on the threshold of extinction. *Sci. Adv.* 5:eaau0757. doi: 10.1126/sciadv.aau0757

Rogers, R. L., and Slatkin, M. (2017). Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet.* 13:e1006601. doi: 10.1371/journal.pgen.1006601

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591. doi: 10.1038/nature08832

Salcedo, T., Geraldes, A., and Nachman, M. W. (2007). Nucleotide variation in wild and inbred mice. *Genetics* 177, 2277–2291. doi: 10.1534/genetics.107.079988

Schultz, S. T., Lynch, M., and Willis, J. H. (1999). Spontaneous deleterious mutation in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 96, 11393–11398. doi: 10.1073/pnas.96.20.11393

Sohail, M., Vakhrusheva, O. A., Sul, J. H., Pulit, S. L., Francioli, L. C., Genome of the Netherlands Consortiumet al. (2017). Negative selection in humans and fruit flies involves synergistic epistasis. *Science* 356, 539–542. doi: 10.1126/science.aah5238

Subramanian, S. (2016). Europeans have a higher proportion of high-frequency deleterious variants than Africans. *Hum. Genet.* 135, 1–7. doi: 10.1007/s00439-015-1604-z

Subramanian, S. (2018). Influence of effective population size on genes under varying levels of selection pressure. *Genome Biol. Evol.* 10, 756–762. doi: 10.1093/gbe/evy047

Subramanian, S. (2021). Deleterious protein-coding variants in diverse cattle breeds of the world. *Genet. Sel. Evol.* 53:80. doi: 10.1186/s12711-021-00674-7

Willemsen, D., Cui, R., Reichard, M., and Valenzano, D. R. (2020). Intra-species differences in population size shape life history and genome evolution. *eLife* 9:e55794. doi: 10.7554/eLife.55794

Xie, X., Yang, Y., Ren, Q., Ding, X., Bao, P., Yan, B., et al. (2018). Accumulation of deleterious mutations in the domestic yak genome. *Anim. Genet.* 49, 384–392. doi: 10.1111/age.12703

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

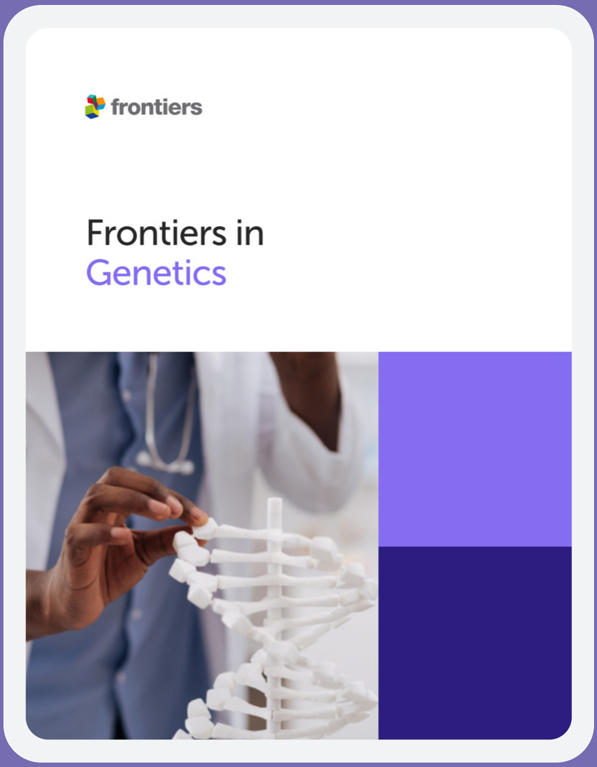
Frontiers in
Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact