Demographic and Selection Histories of Populations Across the Sahel/Savannah Belt

Cesar Fortes-Lima [1], Petr Tříska,[2] Martina Čížková,[2] Eliška Podgorná,[2] Mame Yoro Diallo,[2,3] Carina M. Schlebusch [1,4,5,*] and Viktor Černý[2,3,*]

[1]Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[2]Archaeogenetics Laboratory, Institute of Archaeology of the Czech Academy of Sciences, Prague, Czech Republic
[3]Department of Anthropology and Human Genetics, Faculty of Science, Charles University, Prague, Czech Republic
[4]Palaeo-Research Institute, University of Johannesburg, Johannesburg, South Africa
[5]SciLifeLab, Uppsala, Sweden

***Corresponding authors:** E-mails: cerny@arup.cas.cz; carina.schlebusch@ebc.uu.se.

Associate editor: Dr. Evelyne Heyer

Abstract

The Sahel/Savannah belt harbors diverse populations with different demographic histories and different subsistence patterns. However, populations from this large African region are notably under-represented in genomic research. To investigate the population structure and adaptation history of populations from the Sahel/Savannah space, we generated dense genome-wide genotype data of 327 individuals—comprising 14 ethnolinguistic groups, including 10 previously unsampled populations. Our results highlight fine-scale population structure and complex patterns of admixture, particularly in Fulani groups and Arabic-speaking populations. Among all studied Sahelian populations, only the Rashaayda Arabic-speaking population from eastern Sudan shows a lack of gene flow from African groups, which is consistent with the short history of this population in the African continent. They are recent migrants from Saudi Arabia with evidence of strong genetic isolation during the last few generations and a strong demographic bottleneck. This population also presents a strong selection signal in a genomic region around the *CNR1* gene associated with substance dependence and chronic stress. In Western Sahelian populations, signatures of selection were detected in several other genetic regions, including pathways associated with lactase persistence, immune response, and malaria resistance. Taken together, these findings refine our current knowledge of genetic diversity, population structure, migration, admixture and adaptation of human populations in the Sahel/Savannah belt and contribute to our understanding of human history and health.

***Key words:* Africa, population structure, admixture, selection, pastoralists, camel herders, *CNR1* gene.**

Article

Introduction

The Sahel/Savannah belt, located south of the Sahara Desert, is an important crossroad between northern, western, central, and eastern African groups as well as Eurasian migrants. Human populations from this large region are anthropologically characterized by the co-existence of nomadic pastoralists and sedentary farmers showing their roots in the early Holocene (Černý et al. 2011). Food production in the form of pastoralism started in the Sahara around 8 thousand years ago (ka) (Kuper and Kröpelin 2006). This harsh and inhospitable environment contains relict populations of plants and animals documenting a past environment rich with rivers, freshwater lakes, and grasslands suitable for large herds of ruminants during the early Holocene time periods (Brito et al. 2011; Phelps et al. 2020). Archaeological excavations found evidence of the presence of herders, especially cattle remains, in a great number of archaeological sites in the Sahara (Jesse et al. 2013; Kuper and Riemer 2013). Around 5.5 ka, a rapid climate change in this region led to intensive aridification with a serious impact on all life forms (Manning and Timpson 2014; Armitage et al. 2015). Along with the fauna and flora, human populations gradually emigrated from the newly formed steppe to less hostile environments situated southwards. After this period, the Sahel/Savannah belt became a suitable habitation place lying between the Sahara Desert and the tropical rainforests of sub-Saharan Africa.

Domestication developed independently in the western and eastern Sahelian regions around 4.5 ka (Manning and Timpson 2014; Winchell et al. 2017), and soon after, metallurgy catalyzed further demographic expansions (Maley and Vernet 2015). The emergence of the first African states was also related to social differentiation and accelerated with trans-Saharan trade, where camels and later horses were employed. Due to the constant importance of domestic animals within this semi-desert environment, pastoralists always played a very important role in the cultural history of the Sahel/Savannah belt (McIntosh 2020). A long-term co-existence of ethnically heterogeneous groups of sedentary farmers and nomadic pastoralists was reported in this region (Homewood 2008; Linseele 2013).

Sedentary farmers living in the Sahel are divided into two main groups: extensive and intensive farmers. The extensive farmers, or horticulturalists, cultivate domesticates in a certain time and space but leave once the soil is exhausted. Intensive farmers, or agriculturalists (currently the most numerous group), fertilize the soil, create fields and stay in one place for theoretically unlimited time (Fuller and Hildebrand 2013). Both groups of farmers may keep domestic animals mostly for meat, plowing and fertilizing fields. In contrast, Sahelian nomadic pastoralists represent a specialized group with food-producing strategies adapted to the harsh climate in semi-arid regions. They are specialized in breeding a much higher number of domestic animals per household than neighboring farmers. This requires regular seasonal movements to areas with plenty of pastures (wild grasslands) and water resources. This mode of living is called transhumance and does not provide a possibility to cultivate fields and lead settled lifestyles as intensive farmers do (Pedersen and Benjaminsen 2008; Turner and Schlecht 2019). Therefore, nomadic pastoralists differ from both the extensive and intensive farmers in displaying a much higher degree of mobility.

The population history of the Sahel/Savannah belt has been studied thus far mainly through genetic variation of the uniparental markers (Pereira et al. 2010; Černý et al. 2011; Nováčková et al. 2020; Diallo et al. 2022), or specific genes, such as *LCT*, *NAT2*, *TAS2R16*, or *HLA-B* (Podgorná et al. 2015; Triska et al. 2015; Sanchez-Mazas et al. 2017; Vicente et al. 2019; Priehodová et al. 2020; Kulichová et al. 2021). It has been shown that the genetic differentiation among pastoralists and farmers does not represent significant population structure and that intensive gene-flow or shared ancestry might be among the possible reasons for this (Černý et al. 2021). There is, however, an asymmetry in maternal gene-flow between sedentary farmers and Fulani pastoralists in the western part and between sedentary farmers and Arabic pastoralists in the eastern part of the Sahel belt (Černý et al. 2018). While the Fulani continually loose inherent mitochondrial-DNA (mtDNA) lineages (Fulani women usually marry neighboring farmers), Arab groups, who originally arrived from the Arabian Peninsula, more often accepted women from local sub-Saharan populations into their communities (Čížková et al. 2017). This is also evident from the occurrences of sub-Saharan mtDNA haplotypes in Arabic-speaking populations in the Sahel/Savannah belt (such as the Baggara, Shuwa, and Abbala), while these haplotypes are not found in Arabian populations from founding populations in the Arabian Peninsula (Priehodová et al. 2017).

It has also been shown that genetic distances among local populations of sedentary farmers correlate with geographic but not with linguistic distances (Nováčková et al. 2020). This is more pronounced in analyses of mtDNA data, but occurs also in analysis of Y-chromosome data (Nováčková et al. 2020). In contrast, in pastoralists there is a significant correlation between genetic and linguistic distances but not with geographic distances, for both the maternal and paternal gene pool. These findings suggest that the genetic differentiation among sedentary farmers could be determined by geography, while among nomadic pastoralists the genetic differentiation is influenced by linguistics instead. Interestingly, all sampled local populations of the Fulani pastoralists (collected between Senegal and Chad) have very similar genetic components, and also show genetic affinities with other western African groups. They also have a non-negligible Eurasian contribution (~20%) that is missing in neighboring western African populations (Triska et al. 2015; Vicente et al. 2019; Nováčková et al. 2020; Diallo et al. 2022).

Uniparental markers analyzed among the Sahelian populations thus indicated differences in population history with regards to one or another sex and provided valuable information about gender differences in gene-flow. Uniparental markers however represent only a small portion of the human genome and their low effective sizes and susceptibility to drift effects can distort general population structure patterns in this large African region.

Despite low levels of genetic differentiation, certain genetic variants, such as variants associated with lactase persistence (LP), show large frequency differences between Sahelian populations of different linguistic affiliations or different subsistence forms (Priehodová et al. 2017, 2020; Hollfelder et al. 2021). On the other hand, other genes analyzed in populations from the Sahel/Savannah belt (such as *NAT2* or *TAS2R16*) do not show such a clear differentiation (Podgorná et al. 2015; Kulichová et al. 2021). Genome-wide signals of selection and population structure were previously described in a study focused on a limited number of Sahelian populations (Triska et al. 2015).

To achieve a deeper insight into both the demographic history and genetic patterns of selection in populations from the Sahel/Savannah belt, we generated a genome-wide genotype dataset of Sahelian populations which, together with previously published data, cover the whole Sahel/Savannah belt, spanning from the Atlantic Ocean to the Red Sea (supplementary fig. S1, Supplementary Material online). This dataset included both pastoralist and farmer groups from a broad range of ethnic, cultural, and linguistic groups (including Niger-Congo, Semitic, Cushitic, Omotic, and Nilotic speakers). Our findings highlighted fine-scale population structure between and within regions from the Sahel/Savannah belt, as well as complex patterns of admixture in Fulani and Arabic-speaking populations. Furthermore, we investigated patterns of selection and local adaptation that likely impacted Sahelian populations. We identified putative signals of selection in important genetic regions, including pathways associated with substance dependence and stress in a recent migrant population in Sudan, while signals of selection in immune response genes were found in Western African populations residing in endemic regions for diseases such as malaria.

Results

Genome-wide Diversity and Population Structure in the Sahel/Savannah Belt

To explore population structure within studied Sahelian populations, we first performed principal component analysis (PCA) (Patterson et al. 2006) based on the Sahelian populations genotyped in the present study (supplementary table S1, Supplementary Material online). In the PCA plot (fig. 1*A*), the Rashaayda Arab population from Sudan separate on the first principal component (PC1) from the other studied populations, while on PC2 Western Sahelian populations (in Senegal and Guinea) separate from Central and Eastern Sahelian populations (in Chad and Sudan). We obtained a very similar population structure in PCA plots that display genotyped Sahelian populations and comparative populations across the Sahel/Savannah belt and Yemen from the Sahel-SNP dataset (fig. 1*B* and supplementary fig. S2, Supplementary Material online). Among Sahelian populations, PCA results highlight different clines of genetic variation and suggest admixture in some individuals and populations. On PC1, Arabic-speaking populations are in a cline between the Central and Eastern Sahelian populations and Middle Eastern populations (fig. 1*B*). On PC2, there is another cline of genetic variation between populations in northeastern Africa (e.g., Nuba Koalib) and Niger-Congo-speakers in Western Africa (e.g., Bedik in Senegal). The PC3 (supplementary fig. S2*B*, Supplementary Material online) extremes are defined by the Nuba Koalib population in Sudan, a Kordofanian speaking population, and the Toubou population in Chad, a Nilo-Saharan speaking population. Other populations, such as Bedik in Senegal and Daju in Sudan, split from the remaining populations on PC5 and PC8, respectively (supplementary fig. S2*C* and *D*, Supplementary Material online).

To further explore population structure, we performed PCA based on Sahelian and worldwide populations (fig. 1*C*, supplementary figs. S3–S5 and table S2, Supplementary Material online). PCA results recapitulate genetic differentiation between continental groups with different geographical distributions, linguistic affiliations and lifestyles affiliations (supplementary fig. S4*A*–*D*, Supplementary Material online). On PC2, we detected genetic differentiation between subcontinental groups, for example, Central and Eastern African populations that separate from Western African populations. Among Western Sahelian populations, Fulani groups are between Western African populations and non-sub-Saharan African populations, in a more linear direction to Northern African and Middle Eastern populations than to European populations, suggesting admixture patterns between the former groups and Western African groups as previously reported (Vicente et al. 2019). Among Eastern Sahelian populations, the Rashaayda Arab population has close genetic affinities with Middle Eastern populations like the Yemeni population. On PC3, there is a cline of genetic variation between populations from Western to Eastern Africa. Those three PC projections have significant correlations with geography (supplementary fig. S5, Supplementary Material online), for example, between latitude and PC1 ($r^2 = 0.786$; *P*-value < 1e-5), between longitude and PC2 ($r^2 = 0.493$; *P*-value < 1e-5), and between longitude and PC3 ($r^2 = 0.285$; *P*-value < 1e-5).

We obtained a strong positive correlation between the matrix of genetic distances (pairwise population F_{ST}) and the matrix of geographical distances (Pearson Mantel *r*-statistic = 0.477; *P*-value < 1e-5) (supplementary figs. S6 and S7, Supplementary Material online). As expected, the Rashaayda Arab population has high genetic distances in pairwise comparisons with African populations (supplementary fig. S6, Supplementary Material online). Very strong significant correlations were estimated in comparisons between genetic, linguistic and geographical distance matrices (in all of them the *P*-value was 0.00002), including after controlling for geographical or linguistic diversity (*P*-value = 0.0007 and 0.00002, respectively; supplementary table S3, Supplementary Material online). In general, there were stronger correlations between genetic and geographical distances than between genetic and linguistic distances (Pearson Mantel *r*-statistic = 0.481 and 0.150, respectively). Correlations between genetic and geographical distances after controlling for linguistic diversity (Pearson Mantel *r*-statistic = 0.470, *P*-value = 0.00002) were stronger than correlations between genetic and linguistic distances after controlling for geographical diversity (Pearson Mantel *r*-statistic = 0.0907, *P*-value = 0.00067). The same population structure among Sahelian populations was observed for PCA analyses based on the Low-SNP density dataset, which notably increased the number of reference populations from Eastern African and the Middle Eastern regions (supplementary fig. S8, Supplementary Material online). These results also evidence genetic similarities and possible gene-flow among populations from those regions due to back-migrations to Africa (Fernandes et al. 2019).

To further investigate the population structure and admixture patterns across the Sahel/Savannah belt and worldwide populations, we used clustering analyses as implemented in the ADMIXTURE software package (Alexander et al. 2009), allowing clusters ranging from $K = 2$ to $K = 20$ (supplementary fig. S9, Supplementary Material online). In the ADMIXTURE plot at $K = 4$, the four estimated components are associated with Western African, Central African, Middle Eastern, and European ancestries (fig. 2*A*, *C*, and supplementary figs. S10–S12, Supplementary Material online). In agreement with previous studies (Triska et al. 2015; Hollfelder et al. 2017; Černý et al. 2018), Arabic-speaking populations across the Central and Eastern Sahel have an important genetic contribution from populations with Middle Eastern-related ancestry and complex admixture dynamics with African groups (supplementary table S4, Supplementary Material online), such as the Baggara Arab population residing in Chad and Sudan (on average 27.6% ± 15.6 SD and 20.5% ± 9.1 SD, respectively) and the

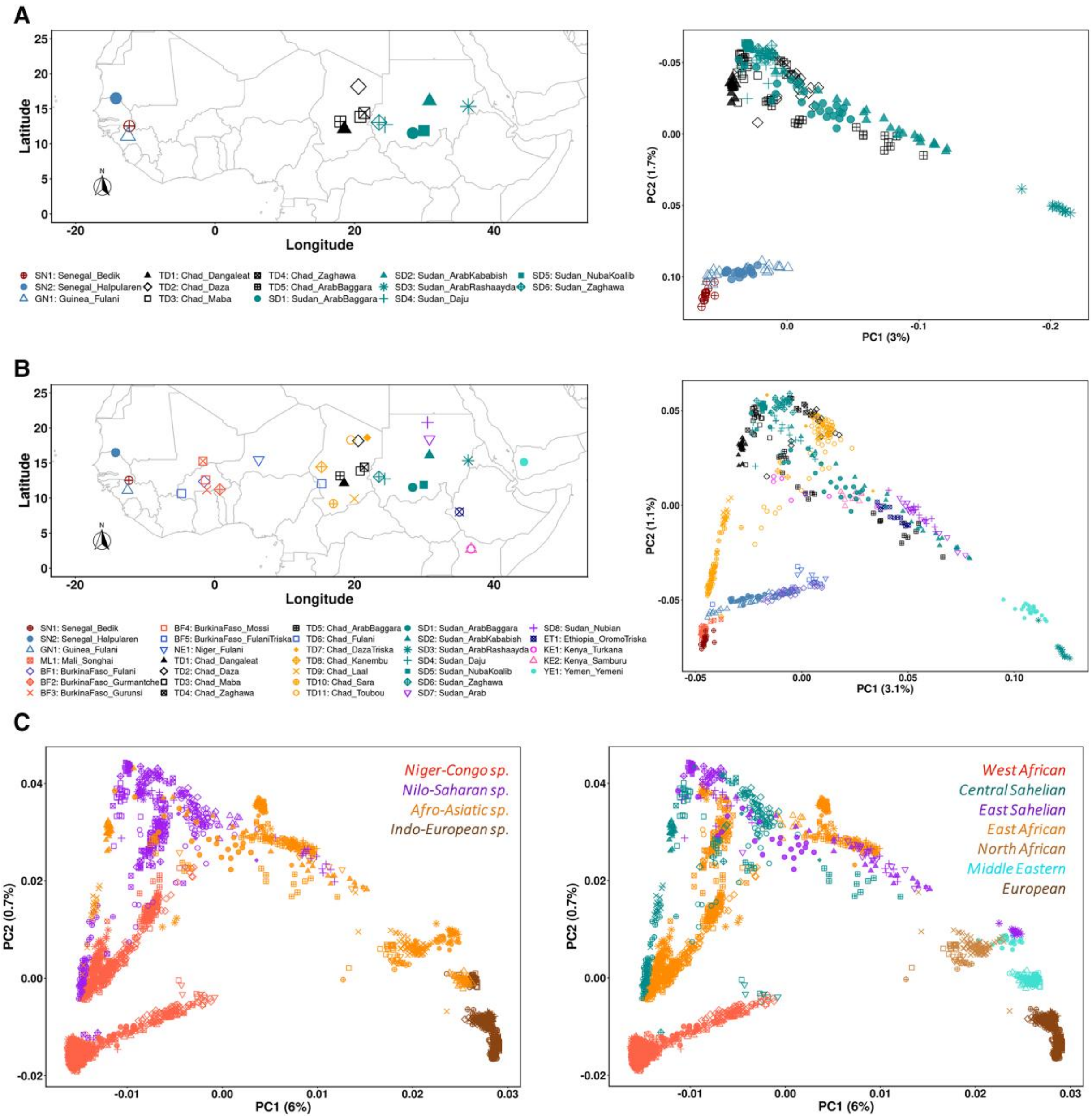

Fig. 1. Geographical locations and PCA for each assembled dataset. (*A*) Locations and PCA results for only Sahelian populations included in the present study. (*B*) Locations and PCA results of Sahelian populations across the Sahel/Savannah belt included in the present study and in previous studies (Triska et al. 2015; Haber et al. 2016; Vicente et al. 2019) (further details were included in supplementary fig. S2, Supplementary Material online). (*C*) PCA plots of worldwide populations included in the High-SNP density dataset (further details were included in supplementary fig. S4, Supplementary Material online). Populations were grouped according to their major linguistic affiliations (left plot) and geographical locations (right plot). Each PCA plot shows the first and second PC projections, and in parentheses the proportion of variance explained for each PC projection. Legend and further details about all the populations included in each dataset are detailed in supplementary fig. S3 and table S2, Supplementary Material online.

Kababish Arab population from Sudan (39.0% ± 16.9 SD). The Rashaayda Arab population from Sudan has the highest values for Middle Eastern-related ancestry (95.1% ± 4.0 SD), which is even higher than in the studied populations from Yemen or Lebanon (ranging from 57.3% ± 1.4 SD to 75.8% ± 4.5 SD).

In ADMIXTURE analysis at $K = 15$, the K-group with the lowest cross-validation (CV) error value (supplementary fig. S12C, Supplementary Material online), the results highlight different genetic components among Sahelian populations (fig. 2*B* and supplementary table S5, Supplementary Material online). For instance, there are two components among Nilo-Saharan-speaking populations, while Western Sahara-speaking populations have high proportions of the teal component (e.g., 79.1% in Toubou and 61.3% in Daza from the northern region in Chad), Eastern Sahara-speaking populations have high proportions of the purple component (e.g., 63.8% in Zaghawa

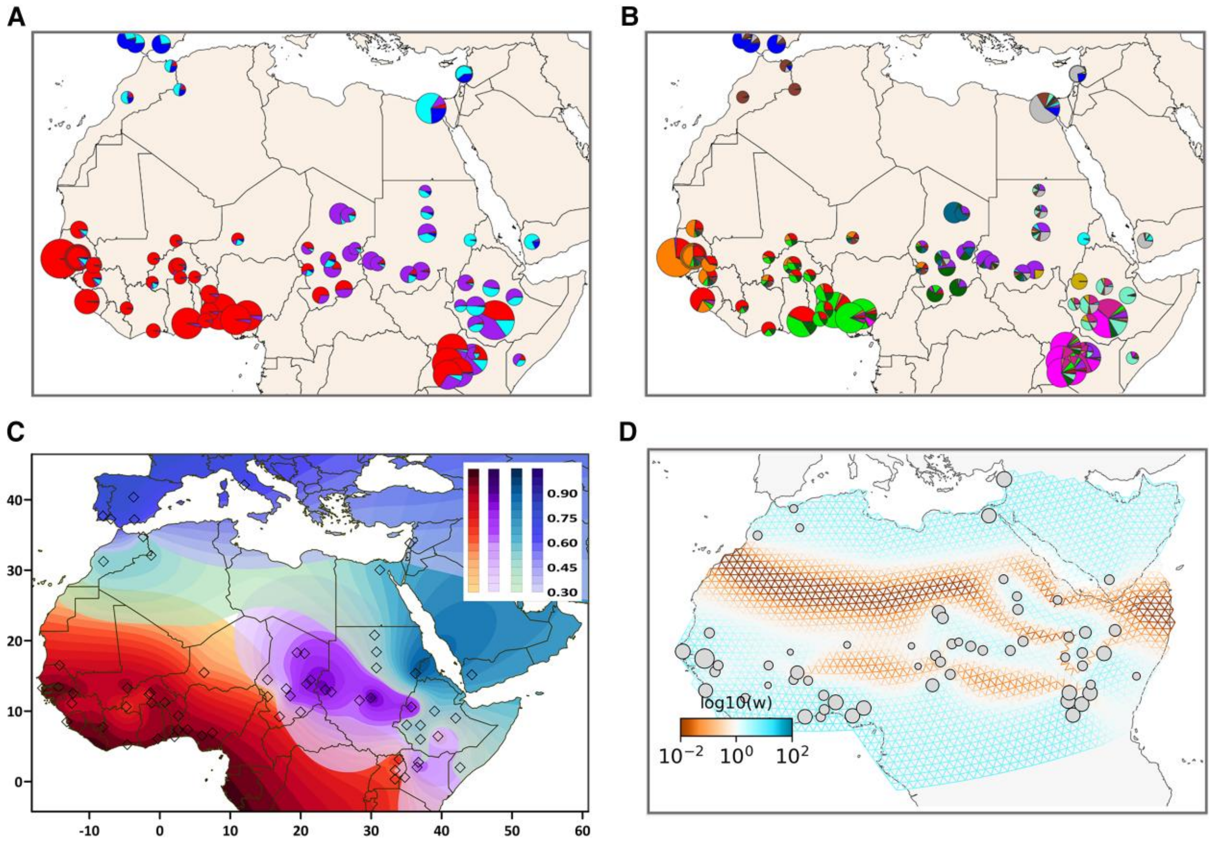

Fig. 2. Patterns of population structure and migration rates across the Sahel/Savannah belt. Figure showing genetic cluster membership estimated for all the populations included in the High-SNP density database using unsupervised clustering ADMIXTURE analysis: (*A*) at $K=4$ and (*B*) at $K=15$ (further details were included in supplementary figs. S9–S13 and tables S4–S5, Supplementary Material online). The size of the pie charts is in relation to the sample size of each studied population. Results for the remaining European populations were included in supplementary fig. S12, Supplementary Material online. (*C*) To better visualize the ADMIXTURE results at $K=4$, the values of the four components were plotted on a geographical map using the Kriging method, highlighting the West African ancestry in red, the Central African ancestry in purple, the Middle Eastern ancestry in light blue, and the European ancestry in dark blue. For each ancestry, cluster assignments below 30% were not included in the figure. (*D*) Effective migration rates estimated using FEEMS. Figure showing fitted parameters in log-scale with lower effective migration shown in orange and higher effective migration shown in blue. A topographic map of the Sahel/Savannah belt was also included (supplementary fig. S10*E*, Supplementary Material online).

in Sudan and 51.8% in Zaghawa in Chad). Interestingly, populations from Chad with different linguistic backgrounds from the southern region in Chad are genetically different from the northern populations. Nilo-Saharan-speaking populations from the northern region in Chad have notably less gene-flow with West-Central African populations than Afro-Asiatic Semitic-speaking populations from the southern region in Chad (supplementary fig. S13, Supplementary Material online). The dark green component is in high proportions in Sara and Laal populations from southern Chad (60.7% and 69.8%, respectively), and is also present in west-central African populations, while the light green has the opposite pattern, suggesting bi-directional gene-flow or shared ancestry between those groups. In Western Africa, studied Fulani populations scattered across five Sahelian countries highlight high proportions of the west and west-central African ancestry from Niger-Congo-speaking populations (range: 30.8–52.2%), as well as admixture from non-sub-Saharan African sources (range: 10.3–29.0%), in agreement with previous genome-wide studies (Busby et al. 2016; Vicente et al. 2019). ADMIXTURE analysis at $K=15$ based on the Low-SNP density dataset (supplementary fig. S14, Supplementary Material online), recapitulates our previous results after increasing the number of Central and Eastern African populations. The results also highlight shared ancestry between the Beja population in Sudan and Afro-Asiatic populations from Ethiopia (Hollfelder et al. 2017).

Kordofanian is a language grouping which, at the moment, is discussed whether it is an isolate or groups with Niger-Congo languages (Quint 2006). A limited number of Koalib-speaking Nuba ($N=16$ individuals) were the only Kordofanian-speaking group that has been studied in genetic studies up until now (Hollfelder et al. 2017). From ADMIXTURE analyses (supplementary figs. S13 and S14, Supplementary Material online), one can see that Nuba populations share more genetic similarities with surrounding populations like the Daju and Dinka (both speaking Eastern Sudanic, Nilo-Saharan languages) rather than Niger-Congo speaking groups from West Africa.

Among all studied populations, the Kordofanian-speaking Nuba Koalib population from the present study has the highest value for the purple component at $K = 15$ (on average 74.4%; fig. 2B, supplementary fig. S13, and table S5, Supplementary Material online). To further investigate the Nuba Koalib population, we performed PCA on the basis of the Only-Sudan database (supplementary fig. S15, Supplementary Material online). PCA of only populations in Sudan highlighted genetic differentiation between several groups in particular the Copts and other populations on PC1, Nuba Koalib and Zaghawa population on PC2, Hausa and other populations on PC3, and Zaghawa and Daju on PC4. The PCA results further highlight genetic relationships between the Nuba Koalib and other populations in Sudan supporting strong genetic affinities with Nilotic populations such as the Dinka and Nuer (supplementary fig. S15, Supplementary Material online). Genetic analyses therefore do not support a link between these two Kordofan-speaking populations and Niger-Congo-speaking populations, rather, they are genetically closest related to surrounding Eastern Sudanic speaking populations.

Migration Rates and Patterns of Shared Haplotypes

To investigate patterns of migration in the Sahel/Savannah belt, we used Fast Estimation of Effective Migration Surfaces software (FEEMS) which allows for depicting spatial population structure and migration surfaces (Petkova et al. 2016; Marcus et al. 2021). The estimated effective migration rates evidenced a very low migration rates between Sahelian and North African populations due to the presence of the geographical barrier represented by the Sahara Desert that limits gene-flow between populations (fig. 2D and supplementary fig. S10F, Supplementary Material online), in agreement with a previous study (Peter et al. 2020). Among Sahelian regions, however, low migration rates were detected also between Western, Central and Eastern groups, likely due to the presence of another geographical barrier represented by Lake Chad, linguistic affiliation, or population history. High migration rates were detected between Nilo-Saharan speaking populations in the Northern part of Chad, Ethiopia and Kenya that highlighted the distribution of Nilo-Saharan groups. We also detected high migration rates between Afro-Asiatic speaking populations in Sudan and Ethiopia.

To infer patterns of shared ancestry between populations across the Sahel/Savannah belt, we estimated sharing patterns of identical-by-descent (IBD) haplotypes (supplementary fig. S16, Supplementary Material online). For short IBD segments (i.e., 3–4 cM), high patterns of sharing were detected between populations from different Sahelian regions suggesting older shared ancestries in each region, for example, between Fulani groups in Western Africa (range: 4.2–17.7 cM per pair); as well as between nomadic populations in Chad (range: 6.1–10.3 cM per pair); and between Arabic-speaking populations in Sudan, for example, the Rashaayda and Kababish (2.96 cM per pair) or between the Rashaayda and Baggara (2.5 cM per pair) (supplementary fig. S16A, Supplementary Material online). In contrast, for long IBD segments (i.e., longer than 5 cM) (supplementary fig. S16B, Supplementary Material online), the Rashaayda Arab population (SD3) shares fewer long IBD segments with nomadic Arabic-speaking populations from the Sahel/Savannah belt. In the northern region of Chad, populations share long IBD segments between them, suggesting they share a common ancestor more recently in their genealogical history, while they share fewer long IBD segments with populations in the southern region of Chad. Fewer patterns of sharing of both short and long IBD segments were detected between Western Sahelian populations and Central or Eastern Sahelian populations, suggesting complex networks of migrations within different regions of the Sahel/Savannah belt.

Homozygosity and Founder Events in the Sahel/Savannah Belt

To shed additional light on the demographic history and cultural practices of Sahelian populations, we analyzed patterns of runs of homozygosity (ROH). Among Sahelian populations, the highest values of the mean number of ROH, total length of ROH, total sum of ROH, and total sum of long ROH longer than 1.5 Mb were observed in Arabic-speaking populations (fig. 3A, B, supplementary figs. S17–S19 and table S6, Supplementary Material online). For the sum of short ROH (<1.5 Mb) (supplementary fig. S18A, Supplementary Material online), the results capture the Out-of-Africa expansion causing Eurasian populations to have higher patterns of homozygosity than African groups (Ceballos et al. 2019). The kurtosis and skewness of the violin plots also provide additional information, while Niger-Congo speaking populations are relatively homogeneous with very short tails and an almost normal distribution, Sahelian Arabic-speaking and Nilo-Saharan populations in Chad and Sudan present more variability with pronounced kurtosis with positive and negative skewness. Interestingly, Fulani groups from the central Sahel region have more homozygosity than Fulani groups from the western Sahel region. In accordance with our previous results, the Rashaayda Arab population has similar values of the sum of short ROH (on average 348.6 ± 21.0 SD; supplementary fig. S18A and table S6, Supplementary Material online) than Middle Eastern populations, which are in-between African and European groups.

Among all populations included in the High-SNP density dataset, the Rashaayda Arab population has the highest values of genomic inbreeding coefficient ($F_{\mathrm{ROH}} = 0.077 \pm 0.033$ SD; supplementary fig. S19A and table S6, Supplementary Material online), which are significantly different (Mann–Whitney *U*-test; *P*-value = 0.00058) from the second population with the highest values (the Kababish Arab population in Sudan (Triska et al. 2015), $F_{\mathrm{ROH}} = 0.038 \pm 0.042$ SD). This population also has the highest values of the total sum of long ROH (230.6 ± 98.6; supplementary fig. S18B, Supplementary Material

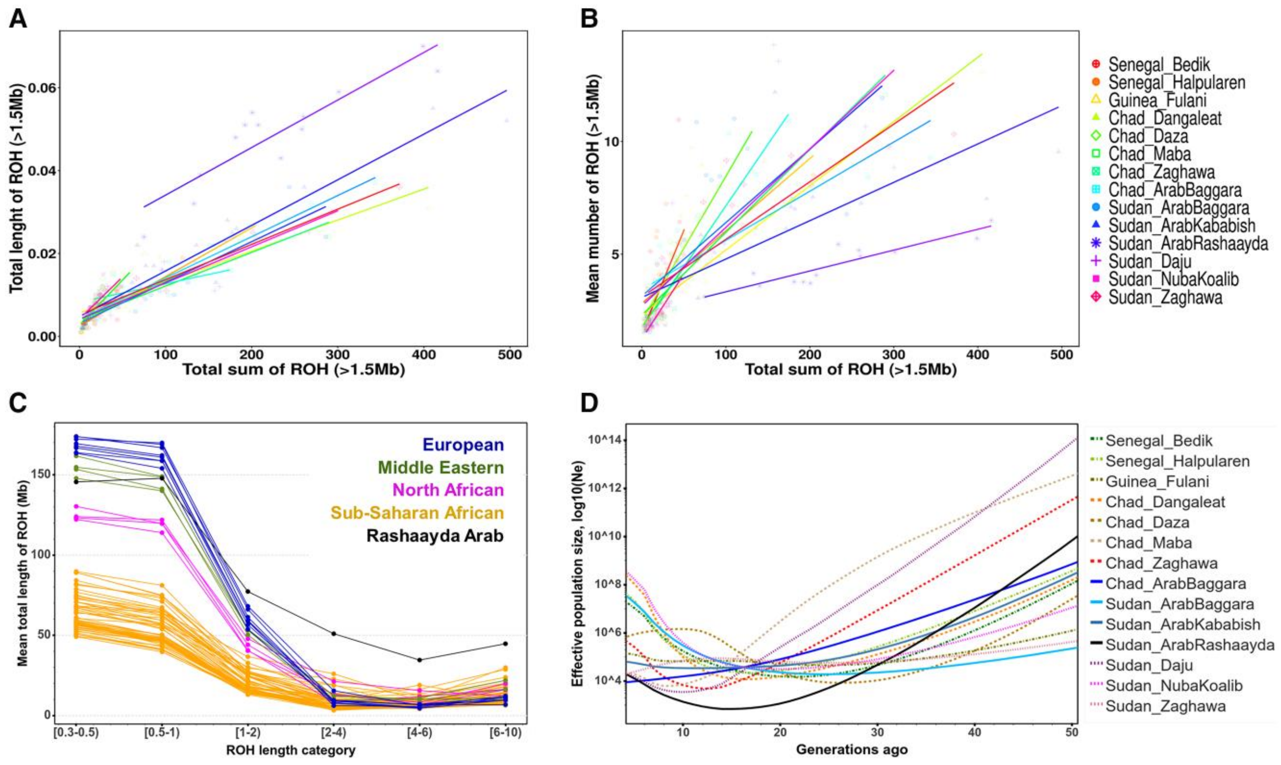

Fig. 3. Patterns of runs of homozygosity (ROH) and effective population sizes in studied Sahelian populations. (*A*) Linear regression between the total length of ROH and total sum of ROH longer than 1.5 Mb in each Sahelian population. (*B*) Linear regression between mean number of ROH and total sum of ROH longer than 1.5 Mb in each Sahelian population. In each figure, linear regressions were estimated using the generalized linear model (GLM) regression function in R. Values of each individual from each population are also shown in the background of each figure (and further details in supplementary fig. S17, Supplementary Material online). (*C*) Mean ROH length categories of all the populations included in the High-SNP density dataset. Figure showing the average for each category for: the Rashaayda Arab population (in black), European populations, Middle Eastern populations, North African populations, and African populations. Further details for each population were included in supplementary figs. S18–S22 and table S6, Supplementary Material online. (*D*) Effective population sizes (N_e) in Sahelian populations for the last 50 generations were estimated using IBDNe (supplementary fig. S25, Supplementary Material online). To better visualize the results, interactive plots were created for figures *C* and *D* (see Interactive_plot-fig_3C.html and Interactive_plot-fig_3D.html).

online) and total length of ROH (0.05 ± 0.01 SD; supplementary fig. S19*B* and table S6, Supplementary Material online). To further investigate patterns of homozygosity, we classified the total sum of ROH into six ROH length classes for each population (fig. 3*C*, supplementary figs. S20–S22 and table S6, Supplementary Material online). For the first and second classes of ROH (i.e., shorter than 2 Mb), high values of ROH were detected in both the Rashaayda Arab population and Middle Eastern populations (range: 125–175 Mb). For the remaining classes of longer ROH (from class 3 to class 6; i.e., from 2 to 10 Mb), the Rashaayda Arab population has the highest averages for those categories, suggesting strong genetic isolation in this population.

Ancestral Origins and Demographic History of the Rashaayda Arab Population

The Rashaayda Arab population living in Eastern Sudan is genetically the most similar to Middle Eastern populations among all the studied Sahelian populations. To investigate the ancestral origins of this population, we used dimensionality reduction methods (i.e., PCA and PCA-UMAP). We first analyzed all the populations from Sudan together with a large representation of Arabic-speaking populations from Africa and the Middle East (supplementary fig. S23, Supplementary Material online), and we then analyzed only the Sudanese Rashaayda together with Middle Eastern populations (supplementary fig. S24, Supplementary Material online). All the results highlight strong genetic affinities between the Rashaayda Arab and Saudi Arabian individuals, in accordance with historical sources (Young 1996).

To investigate the recent demographic changes in Sahelian populations, we estimated effective population sizes (N_e) in the last 50 generations using IBDNe (Browning and Browning 2015). Among all studied populations (fig. 3*D* and supplementary fig. S25, Supplementary Material online), the lowest estimated N_e was detected in the Rashaayda Arab population 15 generations ago ($N_e = 682$; 95% CI = 435–1160; supplementary fig. S25*D*, Supplementary Material online). Also, there is a drop of N_e more or less common for all analyzed populations, with lowest levels around 15 generations ago, and while some populations show recoveries (e.g., Dangaleat in Chad or Bedik in Senegal), others stay demographically unchanged from that time (e.g., Kababish Arab population in Sudan).

In the Rashaayda Arab population, bottleneck events could explain the observed patterns of high homozygosity and low effective population size in the last generations. To infer the age of the founder event and the strength of bottleneck events in Sahelian populations, we performed ASCEND analysis (Tournebize et al. 2022) (supplementary figs. S26, S27 and table S7, Supplementary Material online). Following the four thresholds recommended by Tournebize et al. (2022), we detected significant founder events in the Rashaayda Arab, Baggara Arab, Daju, Nuba Koalib, and Zaghawa populations in Sudan; Dangaleat, Daza, Maba, Zaghawa populations in Chad; and Bedik population in Senegal. The highest estimated bottleneck intensity (I_f) was observed in the Rashaayda Arab population ($I_f = 8.5\%$; 95% CI = 7.9–9.1%; supplementary figs. S26A and S27A, Supplementary Material online), while other studied Sahelian populations have notably lower values of bottleneck intensities since the inferred founder events (I_f range: 0.5–1.8%). As expected (Tournebize et al. 2022), sub-Saharan African populations have lower values than North African and Eurasian populations, except for Fulani populations. We observed high values of bottleneck intensities among Fulani populations that are increasing from western Fulani groups to central Fulani groups (I_f range: 0.3–4.3%). High bottleneck intensity values were estimated in the Songhai population ($I_f = 4.7\%$; 95% CI = 3.2–6.3%), a Nilo-Saharan-speaking population residing in the western Sahel region that has western African ancestry (Triska et al. 2015). The age of the inferred founder event (T_f) in the Rashaayda Arab population was 13 generations ago (95% CI = 11–16) (supplementary fig. S26B, Supplementary Material online), while African and Eurasian populations have older dates (T_f range: 9–194). Among Sahelian populations, the Baggara Arab population in Sudan has the oldest date estimated over 5,000 years ago ($T_f = 179$ generations; 95% CI = 100–258; supplementary table S7, Supplementary Material online), likely this population was founded before the Arab expansion from the Middle East to Africa in the 7th century (Arauna et al. 2017).

Inferring Admixture Events in Populations From the Sahel Belt

To make inferences about admixture events in studied Sahelian populations, we applied different methods. First, admixture f3-statistics tests (Patterson et al. 2012) evidenced significant signals of admixture in Fulani populations from western Africa (supplementary fig. S28A and B, Supplementary Material online), as previously reported (Busby et al. 2016; Vicente et al. 2019). Populations in Sudan such as the Nuba Koalib and Rashaayda Arab lack admixture signals (supplementary fig. S28C and D, Supplementary Material online), as suggested by PCA and ADMIXTURE analyses, in both populations there is null or low level of gene-flow with other groups. Second, MALDER analyses (Pickrell et al. 2014) using multiple reference populations indicated recent admixture times among most of the Sahelian populations (supplementary fig. S29, Supplementary Material online). For all the studied populations, we found a total of 58 significant linkage-disequilibrium (LD) curves weighted by two of the reference populations (supplementary fig. S30, Supplementary Material online). In Sudan, the most recent admixture events were detected in the Baggara Arab population around 16.76 ± 1.84 generations between the Nubian and the Berber Asni in Morocco (supplementary table S8, Supplementary Material online).

Detecting Signatures of Selection

To identify signals that have been targets of selection in populations residing in the Sahel/Savannah belt, we performed genome-wide scan analyses based on three extended haplotype homozygosity (EHH)-based statistics. We first estimated the integrated haplotype score (iHS) across genomic regions in each Sahelian population (supplementary figs. S31–33 and table S9, Supplementary Material online). Among populations from each region of the Sahel/Savannah belt, most of the top 1% candidate regions were private within one of the studied populations rather than between studied populations (supplementary fig. S34, Supplementary Material online), including between Arabic-speaking populations from the same country.

Several signals previously associated with selection in African populations were detected also in Sahelian populations. In the western Sahel region, the Bedik population in Senegal evidenced a strong signal (>4 SD) in chromosome 14 that involved several genes including *SPTB* gene (supplementary fig. S31C, Supplementary Material online, supplementary table S9, Supplementary Material online), which is associated with a rare disorder of the membrane of red blood cells called hereditary spherocytosis (HS) anemia (Mansour-Hendili et al. 2020), and *ACTN1* gene, which is associated to several rare blood disorders (Murphy and Young 2015). Among Fulani individuals sampled in Senegal and Guinea, we detected candidate genomic regions in chromosome 2 around *CTNNA2* gene (supplementary fig. S31E, Supplementary Material online), which is associated with cleft palate in African populations (Butali et al. 2019). In the central Sahelian region, the Zaghawa sampled in Chad has a candidate region near *SLC22A16* gene, which encodes metabolite transporter proteins (Faraji et al. 2016). Interestingly, some Nilo-Saharan-speaking and Arabic-speaking populations shared the same signals for candidate regions of selection. For instance, in both the Daju population in Sudan and the Baggara Arab in Chad (supplementary figs. S32A and S33C, Supplementary Material online), the genomic regions in chromosome 1 displayed significant values (>4 SD) around the *TRABD2B* gene that encodes a cell membrane metalloprotease involved in signaling pathways (Liu et al. 2019), and the *AGBL4* gene, which encode a deglutamylation enzyme (Rogowski et al. 2010).

We then estimated the two pairwise population statistics: cross-population EHH score (XP-EHH) between pairwise populations, and the log-ratio of the integrated site-specific EHH between populations (Rsb) score. XP-EHH scores of the Rashaayda Arab population against several Sahelian and sub-Saharan African and European populations display a strong selection signal in all XP-EHH pairwise comparisons on chromosome 6 (fig. 4A–C, supplementary figs. S35–S37 and table S10, Supplementary Material online). This candidate region is located around the Type 1 Cannabinoid Receptor (*CNR1*) gene (chr6:88,849,584:88,875,767; based on UCSC Genome GRCh37/hg19 [Dreszer et al. 2012]), which is a group of receptors found mainly in the terminals of central and peripheral neurons in the brain (Ashenhurst et al. 2017). In total, 55 single nucleotide polymorphisms (SNPs) from the H3Africa array were included in this inferred candidate region (supplementary table S11, Supplementary Material online). To better visualize XP-EHH results, for each pairwise population comparison, we zoomed in on chromosome 6 and the genomic region of 20 Mb around *CNR1* gene (supplementary figs. S38–S40, Supplementary Material online). We also detected high allele frequencies in two of the top selected SNPs from this candidate region, one SNP in the coding region and another SNP in the transcript region of *CNR1* gene (rs806368 and rs11756397, respectively). Both markers showed notably higher frequencies in the Rashaayda Arab population than in other studied African or Eurasian populations (supplementary figs. S41A, *B* and table S12, Supplementary Material online). Pairwise F_{ST} values between the Rashaayda Arab population and other populations also showed high values with the majority of the studied populations, and the lowest values with the Kanembu for rs806368 and with the Dangaleat for rs11756397 ($F_{ST} = 0.299$ and 0.201, respectively; supplementary table S12, Supplementary Material online), suggesting that those SNPs might also have relatively high values in some African populations from Chad. We also confirmed XP-EHH results after downsampling populations with sample sizes larger than in the Rashaayda Arab population. For the downsampled dataset, allele frequencies showed the same pattern (supplementary fig. S41C and *D*, Supplementary Material online). In agreement, XP-EHH-based scans indicated the same key genomic regions in the pairwise population comparisons after downsampling populations to 13 randomly selected individuals from that population (supplementary figs. S42–S44 and table S13, Supplementary Material online).

In both, XP-EHH comparisons between one Sahelian population and the Rashaayda Arab population or one European population (supplementary tables S10 and S14, Supplementary Material online), several populations from different Sahelian regions (Halpularen, Dangaleat, and Zaghawa) have other candidate regions detected in chromosome 6 for candidate regions that include zinc finger transcription factor genes and protein-coding genes, *ZSCAN12* and *ZKSCAN3*, which are implicated in cancer cell progression and response to bacterial and viral infections (Kanehisa and Goto 2000; Huang et al. 2019; Ouyang et al. 2021).

XP-EHH comparisons between each Sahelian population and one representative sub-Saharan African population with western-central African ancestry (supplementary table S15, Supplementary Material online) evidenced significant values (>3 SD) in candidate regions in chromosome 6 associated with *HLA* genes (e.g., *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1* polymorphism). In the Bedik population from Senegal, the inferred selection region included the HLA class *DQB1*, which is associated with protection against intracellular pathogens such as *Plasmodium vivax* (Lima-Junior and Pratt-Riccio 2016).

In XP-EHH comparisons between each Sahelian population and one representative European population (supplementary table S14, Supplementary Material online), significant selection signals (>3 SD) were detected in the candidate region in chromosome 2 associated with *LCT* gene in the Fulani population sampled in Guinea, in agreement with previous studies of Fulani groups (Triska et al. 2015; Vicente et al. 2019; Cuadros-Espinoza et al. 2022). This candidate region has a derived *LCT* haplotype with a large EHH in studied Fulani individuals (fig. 4*D*).

To confirm the XP-EHH-based test, we also performed Rsb-based pairwise comparisons between the same set of populations. Rsb-based results were in agreement with XP-EHH pairwise comparisons (supplementary figs. S45–S47 and table S16, Supplementary Material online). Likewise, there is a strong signal for selection in pairwise comparisons between other study populations and the Rashaayda Arab population around *CNR1* gene. Across chromosome 6, we observed a strong correlation between comparisons of Rsb *P*-values against XP-EHH *P*-values (supplementary figs. S48-S50, Supplementary Material online). To investigate all the genes from inferred candidate regions of the Rashaayda Arab population against other studied populations, we provided information regarding pathways, gene families, drugs, and diseases of genes that are enriched according to XP-EHH-based scans (supplementary table S17, Supplementary Material online). In total, seven gene families were listed: cannabinoid receptors, topoisomerases, gamma-aminobutyric acid type A receptors, phosphodiesterases, armadillo repeats, leucine zipper proteins, and zinc fingers C2H2-type.

Discussion

The genetic landscape in the Sahel/Savannah belt has been strongly influenced by the geography and demographic history of the populations with a wide range of linguistic affiliations and lifestyles, as has been shown in previous studies (Triska et al. 2015; Kulichová et al. 2017; Nováčková et al. 2020). Our results based on genome-wide genotype data evidence high levels of diversity across this large African region. For all the studied populations, PCA projections have significant correlations with the geographical distribution of studied populations (across both latitude and longitude). This correlation is still significant after correcting for linguistic variation

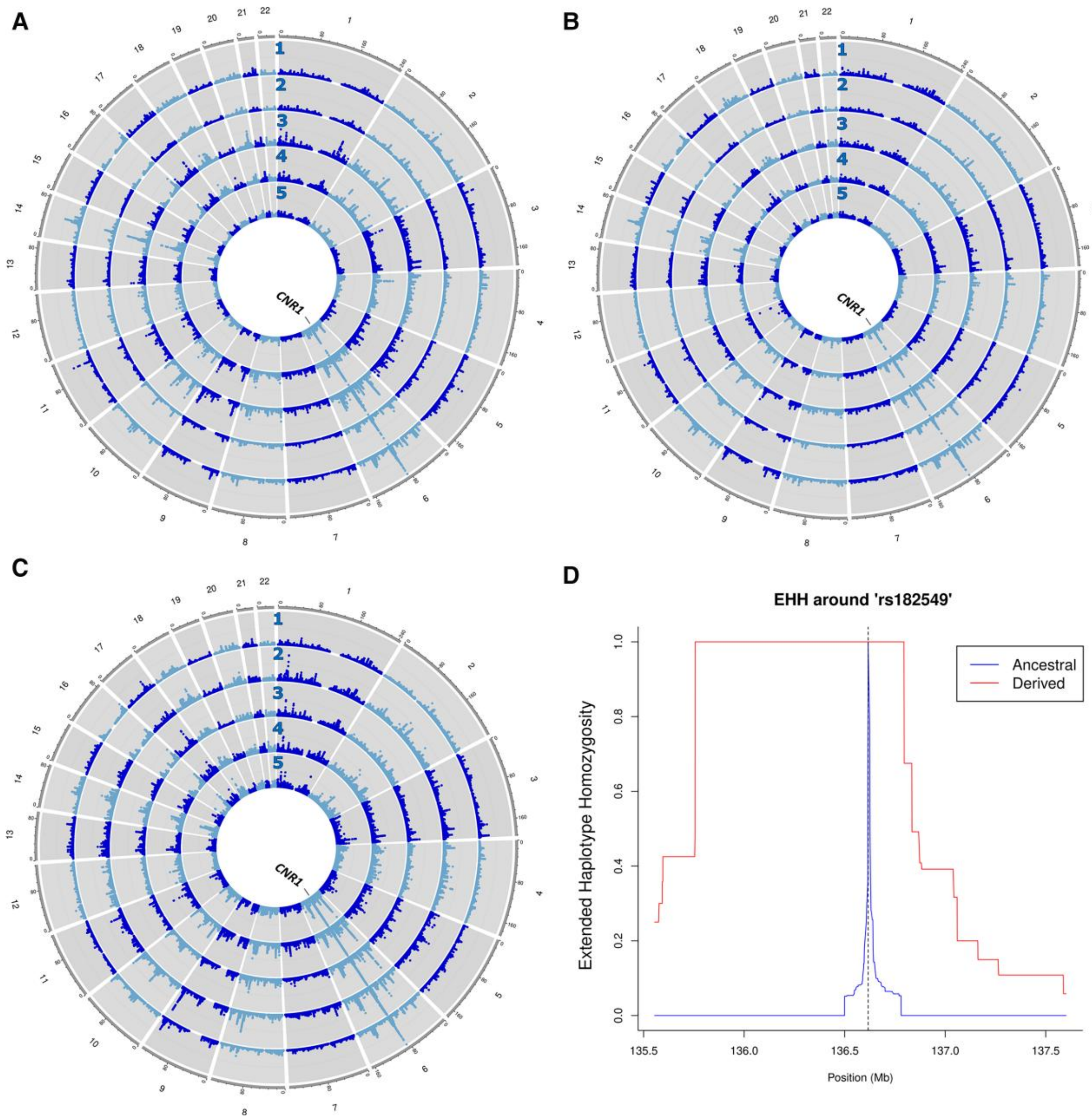

Fig. 4. Genome-wide selection signals in Sahelian populations. XP-EHH-based scan for selection after comparing Rashaayda Arab population and African populations from: (*A*) the Western Sahelian region (1—Gambian in Gambia [GWD], 2—Bedik in Senegal, 3—Halpularen in Senegal, 4—Fulani in Guinea, and 5—Yoruba in Nigeria [YRI]) (supplementary fig. S35, Supplementary Material online); (*B*) the Central Sahelian region in Chad (1—Baggara Arab, 2—Dangaleat, 3—Daza, 4—Maba, and 5—Zaghawa) (supplementary fig. S36, Supplementary Material online); and (*C*) the Eastern Sahelian region in Sudan (1—Baggara Arab, 2—Kababish Arab, 3—Daju, 4—Nuba Koalib, and 5) Zaghawa (supplementary fig. S37, Supplementary Material online). Each figure highlights the location of the candidate region for selection in the *CNR1* gene. Manhattan plots for this chromosome 6 and also a genomic region of 20 Mb around *CNR1* gene were also included (supplementary figs. S38–S40, Supplementary Material online). (*D*) Lactase persistence associated haplotypes in studied Sahelian populations. Figure shows the EHH around the ancestral (in blue) and the derivate (in red) variants, rs182549 (−22018*A), which is in strong LD with rs4988235 (−13910*T), the European lactase prsistence associated variant.

(supplementary table S3, Supplementary Material online). Linguistic affiliations also represent specific patterns and clines of variation in the PCA space, but lifestyle affiliations are more difficult to distinguish between farmers and pastoralist groups (supplementary fig. S4*D*, Supplementary Material online).

Our findings highlight different demographic histories and patterns of migration and admixture in Sahelian groups. These groups have complex social organizations and developed a dense network of migrations across the Sahel/Savannah belt. We detected population structure between and within Sahel/Savannah belt regions, where western Sahelians have notably higher amounts of admixture with Niger-Congo-speaking populations than populations from the other Sahelian regions. In Chad, populations from the northern region have different genetic ancestry components and linguistic backgrounds than populations from the southern region, suggesting that cultural factors or Lake Chad Basin were likely a barrier to population movements within the central Sahelian region (Haber et al. 2016; Magnavita et al. 2019). As expected, a large barrier to human migration was detected in the Sahara Desert (fig. 2*D*), the world's largest desert, which limits migration and gene-flow with north African populations, except for nomadic Fulani groups who show patterns of admixture between western African and north African populations (Vicente et al. 2019). Arabic-speaking populations also evidenced complex patterns of admixture between sub-Saharan African and Middle Eastern populations (Hollfelder et al. 2017; Fernandes et al. 2019). Other Sahelian populations have less admixture and they define different clusters suggesting a high genetic differentiation. One cluster was defined by the Toubou, a Nilo-Saharan-speaking population from Chad, while the another cluster was defined by the Nuba Koalib, a Kordofanian-speaking population from Sudan (supplementary figs. S2, S13 and S14, Supplementary Material online). Although speaking a different language family (Kordofanian) with uncertain and debated connections to Niger-Congo, the Nuba Koalib population has genetic similarities with other populations from Sudan, such as the Dinka and Daju who speak Nilo-Saharan languages, rather than with Niger-Congo speakers. This is also in accordance with the results based on uniparental markers (Nováčková et al. 2020).

The Rashaayda Arab population represents nomadic Arabic-speaking people who practice camel-herding. They currently live in scattered areas predominantly in Eritrea and North-Eastern Sudan. Among all studied Sahelian populations, this recent migrant population has a strong shared ancestry with the Middle Eastern groups and no or very little gene-flow with African groups. In agreement, the Rashaayda Arab population has mtDNA haplogroups mostly present among only Middle Eastern populations, such as R0a2c and J1b (Čížková et al. 2017; Priehodová et al. 2017), and in all genetic results are striking outliers among African populations due to genetic drift or isolation. We found strong genetic affinities between this population and Arabic-speaking people living in Saudi Arabia (supplementary figs. S23 and S24, Supplementary Material online), highlighting their common ancestral origins. In contrast to other Middle Eastern populations from previous studies (Haber et al. 2016; Fernandes et al. 2019), the Rashaayda Arab population has very low levels of genetic admixture with African groups (supplementary fig. S28D, Supplementary Material online). This is likely influenced by the fact that Saudi Arabia received notably less gene-flow from sub-Saharan African groups compared with other Middle Eastern populations such as the populations in Yemen (Abu-Amero et al. 2007; Černý et al. 2008, 2016; Fernandes et al. 2019). In accordance, historical sources point out that the Rashaayda is a Bedouin group, descending from Banu Abs, an indigenous group in the Hejaz region of Saudi Arabia (Young 1996). They are also known as Rashaida or Bani Rashid people and migrated from Saudi Arabia to Eritrea and Sudan in the mid to late 1800s (Young 1996).

This recent migrant population has been in strong genetic isolation likely due to cultural barriers and shows a high prevalence of consanguineous marriages. Their elevated patterns of homozygosity and the timing of the estimated founder event suggest strong patterns of genetic isolation that even predate their arrival in Sudan around two centuries ago. Therefore, this population must have been a minority group in Saudi Arabia several generations before they migrated to Eastern Sudan. Historical sources highlight that the Rashaayda was a marginalized group in Saudi Arabia due to ethnic warfare and starvation (Young 1996). Our results show that the Rashaayda Arab population decreased around 15 generations ago (~1515 CE) in the effective population size (supplementary fig. S25*D*, Supplementary Material online), likely when they were still in the Arabian Peninsula. Thereafter, this population migrated to Sudan, where they experienced a strong founder event that further increased their patterns of homozygosity (fig. 3C and supplementary fig. S26, Supplementary Material online). According to anthropological studies, marriages in the Rashaayda Arab population usually follow Arabic cultural traditions and are arranged by families (Young 1996). In nomadic Sahelian Arabic-speaking societies, there is a rule in the choice of partners, according to which a girl should marry a parallel cousin on her father's side, that is, the son of her father's brother, if available (Cunnison 1966; Asad 1970).

Due to the hard environment in the Sahel/Savannah belt, several droughts and famine affected essentially all the populations resigning in this region (Carré et al. 2019). We detected the lowest levels in effective population sizes around 15 generations ago (approximately in the 17th century AD). At that time the climatic conditions do not seem to have been the main cause of population decline, for instance Lake Chad experienced a rise in water level right at this time (Brunk and Gronenbom 2004). In addition to climatic factors, cultural factors might also have affected effective population sizes, for example the variance of reproductive success. Since 10th century AD, Sahelian populations have been under the influence of Islamization by the Arab populations, who practice endogamy and consanguineous matrimonial practices that can reduce effective population sizes, as can be seen in the Arab Rashaayda population (Young 1996). However, these interpretations should be viewed with caution, since information on marital practices in earlier generations of the Sahelian populations is not available.

The only non-Muslim population in our dataset was the Bedik from southeastern Senegal. This population, together with the Dangaleat population from northern Chad, show differences in the pattern of change in N_e (drop and recovery) from other studied Sahelian populations (fig. 3*D*). The Dangaleat belong to the group of people called the Hadjeray (from the Arabic "those of the stones"), whose ancestors resisted raids (slave hunting) organized by the Wadai Empire in the 17th century. During this time period, they still were non-Muslims, and their conversions were more recent in time.

From a wider perspective, it is difficult to classify Sahelian populations as endogamic or exogamic to assess this effect on effective population sizes. Endogamy is present in most Sahelian populations; the only difference might be the proportion of endogamous partnerships in a population. For example, in the Fulani pastoralists from Burkina Faso, the ratio of such unions is relatively high (65.8% for women and 71.0% for men) on average (Hampshire and Smith 2001). However, it has been also shown that higher rates of endogamy occur in places where there is not enough pasture for larger herds of cattle; especially evident among the poorest pastoralists, who have less than five cows at their disposal (Hampshire and Smith 2001). Therefore, the rates of endogamy might also depend on the locations from where the samples were obtained for the study.

The genetic diversity influencing the LP trait has been studied in Sahelian populations (Ranciaro et al. 2014; Hollfelder et al. 2021). Strong evidence of a founder event in the Rashaayda population was inferred in previous studies of the -13915*G *LCT* variant (Priehodová et al. 2014, 2017). This variant is associated with camel herders from the Middle East (Enattah et al. 2008; Priehodová et al. 2017), and its high frequencies in those populations are notably different from populations in Sudan and South Sudan (ranging between 0% and 28%) (Hollfelder et al. 2021). However, the highest frequency for this *LCT* variant (on average 76.5% in 51 individuals) and the highest predicted LP phenotype (95%) have been found in the Rashaayda population residing in the eastern Sahelian region (Priehodová et al. 2020). In Fulani groups, we detected a large haplotype around the European *LCT* -22018*A (rs182549), which is in strong linkage disequilibrium (LD) with −13910*T (rs4988235), the European *LCT* variant (Enattah et al. 2002). This haplotype was likely introduced to the western Sahel via admixture with a northern African population with Eurasian ancestry (Vicente et al. 2019). However, further research has shown that this variant is not present in high frequencies only in populations of the Fulani pastoralists, but also in other western Sahelian pastoralist such as the Tuareg and Moors (Priehodová et al. 2020).

A specific signal of exceptionally strong selection involving the *CNR1* gene in the Rashaayda Arab population has been highlighted in this study. In all pairwise population comparisons, both XP-EHH- and Rsb-based genome-wide scans gave consistent results. *CNR1* regulates both the endocannabinoid and dopaminergic neurobiological systems, and polymorphisms of this gene have been previously associated with substance dependence, such as morphine or cocaine (Clarke et al. 2013), as well as regulation of neuronal and endocrine responses to chronic social defeat stress (or CSDS) due to depression or anxiety-like behaviors (Beins et al. 2021). The associations of mutations in *CNR1* gene with the risk of developing substance dependence have been recently demonstrated (Pabalan et al. 2021). In the Rashaayda population, the *CNR1* gene could likely be under selection due to cultural adaptations of this population due to chronic stress (e.g., CSDS) or substance dependence (e.g., use of morphine obtained from poppy tears). However, stress has a pleiotropic nature controlled by numerous genes and likely other genes and functional pathways might be involved as well. Future integrative genomic research will shed light on the functional impacts of variants under selection for adaptation to environmental stress or lifestyle factors. In addition, future association studies of molecular and clinical phenotypes would clarify the context-specific effects of this region/locus and its impact on downstream phenotypes.

In Western African populations residing in malaria-endemic regions, we detected candidate regions of selection that are associated with malaria genes, in particular *SPTB* and *ACTN1* genes and *HLA* polymorphisms. Several populations from different Sahelian regions have candidate regions that include zinc finger transcription factor genes (*ZSCAN12* and *ZKSCAN3* genes), which are implicated in response to cancer progression and bacterial and viral infections.

In summary, the Sahel/Savannah belt harbors diverse populations with different demographic histories and different genetic backgrounds. Our results evaluated the level of population structure and patterns of admixture, particularly among Fulani groups from the western part of the Sahel and Arabic-speaking populations from its eastern part. In contrast to other Arabic-speaking populations, the Rashaayda Arabs in Sudan show a lack of gene-flow with any sub-Saharan African population, which is consistent with their recent arrival and short history in the African continent. Rashaayda Arabs in eastern Sudan are recent migrants from Saudi Arabia with high values of inbreeding coefficients and patterns of homozygosity, evidencing strong genetic isolation during the last generations, and a bottleneck around 15 generations ago. This population has a strong selection signal in a genomic region around the *CNR1* gene associated with substance dependence and CSDS. In Western African populations, signatures of selection were detected in several genetic regions, including pathways associated with the LP trait and immune responses triggered by endemic diseases such as malaria. Taken together, these findings indicate complex evolutionary histories in Sahelian populations from different regions, shedding new light on how selection may have influenced their adaptation to harsh climate and environmental changes in semi-arid Sahelian regions.

Materials and Methods

Sampling Design

We collected saliva samples of populations from three regions of the Sahel/Savannah belt the: Western region (three populations in Senegal and Guinea), Central region (five populations in Chad), and Eastern region (six populations Sudan) (supplementary fig. S1, Supplementary Material online). Selected populations from different geographical locations represent groups with different subsistence strategies (farmers, pastoralists and agro-pastoralists) and belong to three linguistic affiliations (Niger-Congo, Nilo-Saharan, and Afro-Asiatic linguistic families) (supplementary tables S1 and S2, Supplementary Material online). Each participant gave informed consent before donating their saliva sample, and only individuals whose parents and grandparents came from the same ethnolinguistic group were included. This study was approved by the Ethical Committee of the Charles University in Prague (the Czech Republic, approval no. 2019/12) and the Swedish Ethical Review Authority under the Ministry of Education (Sweden, approval no. 2 2019-00479) and was conducted according to the Declaration of Helsinki.

Genotyping Procedure

DNA was extracted from each saliva sample following the protocol provided by the supplier (DNA Genotek Inc.). For this study, we generated genome-wide SNP data of 327 participants in total from 14 Sahelian populations, on average 23 participants per population (ranging from 12 to 25 individuals per population) (supplementary table S1, Supplementary Material online). DNA samples were genotyped on the Illumina H3Africa array (BeadChip type: H3Africa_2017_20021485_A2; designed for SNP-genotyping of 2,267,346 SNPs) at the SNP&SEQ Technology Platform, NGI/SciLifeLab Genomics (Sweden). This genotyping array was designed to account for the larger genetic diversity and smaller haplotype segments in African populations (Mulder et al. 2018). After genotyping, the average SNP call rate per individual was 99.4% (ranging from 83.5% to 99.7%).

Assembling Genome-wide Genotype Datasets

For autosomal data, quality control (QC) steps were performed using PLINK v1.9 (Chang et al. 2015), to keep autosomal biallelic variants with a high genotyping rate (plink --mind 0.15 --geno 0.1 --hwe 0.0000001). Three samples were removed due to their low genotyping rate (mind > 0.15). To estimate recent genetic relatedness, we used KING (Manichaikul et al. 2010) and PC-Relate analysis from GENESIS software (Conomos et al. 2016), and we removed 59 individuals due to first- or second-degree kinship (supplementary fig. S51, Supplementary Material online). For our new dataset, we obtained 2,206,620 SNPs and 268 individuals. We also used PLINK to prune SNPs under high LD (plink --indep-pairwise 100 10 0.2) for analyses that assume unlinked variation and the LD-pruned dataset contains 459,880 variants and 268 individuals (supplementary table S1, Supplementary Material online).

To investigate genetic affinities in our dataset, we built three genome-wide SNP datasets for different types of analyses (see details of populations included in each dataset in supplementary table S2, Supplementary Material online). First, we assembled the “Sahel-SNP” dataset by merging the QC-filtered dataset from this study with selected populations from the Sahel/Savannah belt presented in previous genome-wide studies using the Illumina HumanOmni2.5 array (Triska et al. 2015; Haber et al. 2016; Vicente et al. 2019). Before and after each merging, we applied the same QC steps using plink. After merging, we obtained 651 Sahelian individuals from 30 populations and 1,229,657 SNPs for the “Sahel-SNP” dataset (and 360,282 SNPs for the Sahel-SNP LD-pruned dataset).

Second, we assembled the “High-SNP density” dataset by merging the QC-filtered Sahel-SNP dataset with datasets from worldwide populations included in previous studies (1000 Genomes Project Consortium et al. 2015; Gurdasani et al. 2015; Haber et al. 2016; Fortes-Lima et al. 2017; Hernández et al. 2020) (supplementary table S2 and fig. S1, Supplementary Material online). After merging and QC-steps, the High-SNP density dataset contains 2,967 individuals from 94 populations and 813,052 SNPs (and 258,994 SNPs for the "High-SNP LD-pruned dataset").

Third, to investigate the ancestral origins of migrant populations in the Eastern region of the Sahel/Savannah belt, we assembled a dataset with Eastern African and Middle Eastern populations, we assembled the “Low-SNP density” dataset by merging the High-SNP density dataset and publicly available datasets from previous studies (Hollfelder et al. 2017; Fernandes et al. 2019; Scheinfeldt et al. 2019) (supplementary table S2 and fig. S1, Supplementary Material online). After merging and QC-steps, the Low-SNP dataset contains 4,321 individuals from 139 populations and 181,004 SNPs (and 110,220 for the "Low-SNP LD-pruned" dataset). We also created a subset database for populations in Sudan from the present study (excluding the recent migrant Rashaayda Arab population) and from previous publications (Triska et al. 2015; Hollfelder et al. 2017), called “Only-Sudan” dataset (containing 370 individuals from 25 populations in Sudan and 472,291 SNPs).

Correlations Between Linguistic, Geographical and Genetic Distances

To test the correspondence between language, geography, and genetic diversity, we performed a Mantel test using the R package *ncf*. For genetic distances, we used smartPCA tool from the EIGENSOFT package (Patterson et al. 2006) to calculate pairwise Hudson F_{ST} between populations, which produces estimates that are independent of sample sizes in uneven populations (Bhatia et al. 2013). Geographic distances were calculated as great circle distances (accounting for the curvature of the earth) using the R package *fields* (Nychka et al. 2005). For linguistic distances, we estimated distances between pairs of languages using the Levenshtein distance normalized divided algorithm defined by Bakker et al. (2009). To do so, we applied the automated similarity judgment program (Wichmann et al. 2020), which estimates lexical distances by comparing 100-item word lists of all the languages spoken by the populations included in the High-SNP density dataset (supplementary table S18, Supplementary Material online). We first estimated correlations between geographical and genetic distances using Mantel test (Mantel 1967) based on Pearson's *r*-statistic, and significance was estimated using 100,000 random permutations. We then estimated correlations between geographical, linguistic, and genetic distances using partial Mantel tests.

Population Structure Analyses

To investigate the population structure across populations from the Sahel/Savannah belt, we first performed PCA using *smartpca* (Patterson et al. 2012) for the new data and the three assembled datasets. We plotted the results using PCAviz (Novembre et al. 2019; Liu et al. 2020), and in-house R scripts. To combine the information of the first ten PCs, we then performed PCA-UMAP (Diaz-Papkovich et al. 2019) for a subset of selected populations from Eastern Africa and the Middle East.

For the High-SNP and Low-SNP LD-pruned datasets, we used ADMIXTURE software v1.3 (Alexander et al. 2009; Alexander and Lange 2011) to carry out unsupervised ADMIXTURE analysis for each dataset from $K = 2$ to $K = 20$, and for 10 independent runs with a random seed for each K-group. The CV test was performed for each run of each K-group (supplementary fig. S12C, Supplementary Material online). ADMIXTURE analysis was carried out using the projection mode for the admixed Fulani groups included in the High-SNP density dataset, where Fulani samples were projected onto the population structure (allele frequencies) learned from the reference panels included in the dataset. To better visualize the major mode of the ADMIXTURE results, we used AncestryPainter graphic program (Feng et al. 2018). For a better comparison, the width of each population was set to be equal regardless of its sample size. To investigate the geographical distributions of ADMIXTURE results, we plotted estimated admixture proportions on a geographic map. We used the grid-based mapping Surfer software (Golden Software, LLC) and applied the Kriging method for spatial interpolation.

Estimating Effective Migration Patterns

To further investigate spatial population structure in the Sahel/Savannah belt, we used FEEMS (Petkova et al. 2016; Marcus et al. 2021). Briefly, FEEMS applies a Gaussian Markov random field model in a penalized-likelihood-based framework to infer whether populations are exchanging gene flow with neighboring populations in a spatial graph of a "stepping-stone" model of migration and genetic drift. To estimate effective migration parameters, we used as inputs the genotype data of African and Middle Eastern populations included in the High-SNP density dataset, as well as the latitude and longitude coordinates of each population. This analysis is suitable for the dataset that we collected in geographically different sites/locations, in local villages, camps etc. and not big cities or hospitals, making the samples more representative of local demes in the Sahel/Savannah belt.

ROH and Genomic Inbreeding Coefficients

To calculate ROH, we used a sliding-window approach implemented in PLINK v1.9 following recommendations from a previous study estimating ROH in African populations (Ceballos et al. 2019). First, we calculated ROH for each population included in the High-density SNP dataset. ROH were detected with the following parameters: 30 was the minimum number of SNPs that a ROH was required to have (--homozyg-snp 30), 300 was the length in kb of the sliding window (--homozyg-kb 300), 30 was the required minimum density to consider a ROH that means 1 SNP in each 30 kb (--homozyg-density 30), 30 was the number of SNPs that the sliding window must have (--homozyg-window-snp 30), 1,000 was the length in kb between two SNPs to be considered in two different segments (--homozyg-gap 1000), 1 was the number of heterozygous SNPs allowed in each window (--homozyg-window-het 1), 5 was the number of missing calls allowed in a window (--homozyg-window-missing 5), and 0.05 was the proportion of overlapping window that must be called homozygous to define a given SNP as in a "homozygous" segment (--homozyg-window-threshold 0.05). For ROH longer than 1.5 Mb, we calculated mean ROH size, sum of long ROH, and total length of ROH; and for ROH shorter than 1.5 Mb, we calculated the sum of short ROH; using an available R script (https://github.com/CeballosGene/ROH). The genomic inbreeding coefficient based on ROH (or F_{ROH}) measures the actual proportion of the autosomal genome that is autozygous. We estimated F_{ROH} as the total sum of ROH >1.5 Mb divided by the total length of the autosomal genome (3 Gb) (McQuillan et al. 2012; Ceballos et al. 2018). We plotted the results using custom R scripts. Linear regressions between ROH parameters in each population were then estimated using the generalized linear model (GLM) regression function in R. We also calculated six ROH length classes: class 1 ($0.3 < \text{ROH} < 0.5$ Mb), class 2 ($0.5 < \text{ROH} < 1$ Mb), class 3 ($1 < \text{ROH} < 2$ Mb), class 4 ($2 < \text{ROH} < 4$ Mb), class 5 ($4 < \text{ROH} < 6$ Mb), and class 6 ($6 \text{ Mb} < \text{ROH} < 10$ Mb). For all the ROH length classes together, we used an in-house Pythin script with the bokeh visualisation library for interactive plots in ".html" format, with the option of scroll-over metadata for better identifying on the plot the results of each studied population.

Estimating the Timing of Founder Events

To infer both the age and strength of demographic founder events in Sahelian populations, we used ASCEND v8.6 (Tournebize et al. 2022). This method uses the correlation in allele sharing across the genome between pairs of individuals to recover signatures of past bottlenecks in each studied population. For each population included in the High-SNP density dataset, we used default settings for all the autosomal chromosomes, such as distance bins from 0.1 cM to 30.0 cM by steps of 0.1 cM. We estimated the age since the founder event (*Tf*, in generations before the present) and the intensity of the founder event (*If*). To evaluate the quality of the exponential fit, we estimated the normalized root-mean-square deviation between the empirical allele sharing correlation values and the fitted ones, and we plotted the correlations between empirical and theoretical decay curves. To identify significant founder events, we followed the four criteria recommended by Tournebize et al. (2022). To convert the inferred dates since the founder event from generations to years, we used the following equation: 1950-(g*29), where g is the estimated number of generations and 29 is the assumed number of years for one generation (Fenner 2005).

Sharing of IBD Segments

To infer population dynamics, we analyzed identity-by-descent (IBD) segments for selected populations included in the High-SNP dataset. We employed the fastIBD algorithm in the Beagle package (Browning and Browning 2011) to first calculate the total amount of shared IBD segments between individuals in tested populations. We then calculated the average amount of shared IBD in centimorgans (cM) between individuals in tested populations. We removed IBD segments of 3 cM to avoid the conflation effect of short IBD segments (Chiang et al. 2016), and parameters for IBD sharing were calculated separately for short IBD segments (3–4 cM) and long IBD segments (5 cM or more). Because the length of IBD segments decreases over generations due to recombination, short IBD segments suggest ancient common ancestors of two populations, while long IBD segments suggest evidence of a recent gene-flow between pairwise populations. To estimate the effective population size (N_e) in the last 50 generations of each studied Sahelian population, we analyzed IBD-segment data using IBDNe (Browning and Browning 2015) with a threshold size of 2 cM. For IBDNe plots, we included an interactive plot in ".html" format with the option of scroll-over metadata for better identifying the results of each population on the plot.

Admixture Inference Methods

To confirm admixture results, we tested for admixture using f3-statistics (Patterson et al. 2012) in the form f3(A; B, C). If the test results are positive (i.e., f3(A; B, C) > 0), then there is no evidence that target population A is descended from an admixture event between population source B and C. If the test results are significantly negative (i.e., f3(A; B, C) > 0), then target population A might have the results of admixture between population source B and C (Lipson 2020). To make inferences about admixture times in studied Sahelian populations, we calculated the weighted LD statistic using the implementation of ALDER (Loh et al. 2013) for multiple events of admixture called MALDER.

Genome-wide Scan Analyses for Selection

To identify recent signatures of positive selective sweeps, we perform haplotype-based selection scan analysis using rehh v3.2.2 (Gautier and Vitalis 2012; Gautier et al. 2017). We estimated the extent of haplotype homozygosity (EHH) within studied Sahelian populations using the integrated haplotype homozygosity (iHS) score (Voight et al. 2006) and between populations using the cross-population EHH (XP-EHH) score (Sabeti et al. 2007) and the log-ratio of the integrated site-specific EHH between pairwise populations (Rsb) score (Tang et al. 2007). We phased the High-SNP density dataset using SHAPEIT v2 (Delaneau et al. 2013) with the Haplotype Reference Consortium as a reference panel (Consortium and the Haplotype Reference Consortium 2016), and recombination maps were interpolated from the HapMap Phase 2 genetic map. To minimize switch error rates (Delaneau et al. 2013), we used the following parameters: 500 states, 50 MCMC main steps, 10 burnin and 10 pruning steps. First, we calculated the iHS score in each population for all the variants with a minor allele frequency higher than 0.05 to exclude variants near to fixation, and also variants were excluded from the calculation if a 20 kb maximum gap between two SNPs was found when calculating EHH, as they may produce biases. Then, we estimated XP-EHH between pairwise populations, using one studied Sahelian population versus one reference population for western African, European and Middle Eastern ancestries (YRI, CEU and Rashaayda, respectively). To confirm XP-EHH-based tests, we also estimated Rsb statistics between each studied population and the Rashaayda Arab population. For each SNP in each population, we computed a weighted average of the EHH at both alleles, referred to as site-specific EHH (EHHS). Rsb scores were then calculated for the observed distribution of the standardized log-ratio of the integrated EHHS (iES) between pairs of populations, and genomic regions with unusually high Rsb evidenced signals of positive selection. Also, we compared Rsb-based *P*-values against XP-EHH-based *P*-values across chromosome 6.

To identify the ancestral alleles in our dataset, we only analyzed SNPs that are present in the genomes of the three great apes (chimpanzee, orangutan and gorilla). For each 30 kb window of iHS, XP-EHH and Rsb score, we calculated the mean, maximum and *P*-value values. For each statistic, all estimated *P*-values were adjusted for multiple comparisons by using the Benjamini & Hochberg correction algorithm (Benjamini and Hochberg 1995), which is recommended to reduce the false discovery rate (François et al. 2016). Each score was constructed to have an approximately standard Gaussian distribution and be transformed into a *P*-value (in a −log10 scale), and plotted using the Manhattan-plot function in rehh (Gautier et al. 2017). As a genome-wide threshold, we used the quartile 1% for the scores of each population, and candidate regions for selection were inferred for all the genetic windows above that threshold. To annotate the gene-coding content of each candidate region, we intersected windows with the hg19 gene annotations from the “RefFat.txt” file included in the UCSC genome annotation database (Dreszer et al. 2012). For pathway enrichment analysis, we used ToppGene software (Chen et al. 2009) to list pathways, gene families, drugs and diseases of all the genes that were enriched according to XP-EHH-based scans (supplementary table S17, Supplementary Material online).

XP-EHH results were plotted in rounded plots using the R package *shinyCircos* (Yu et al. 2018). To visualize results on chromosome 6, we plotted in Manhattan-plots and we also zoomed into a particular genomic region of 20 Mb, and plotted allele frequencies in African and Eurasian populations for two selected SNPs from that region. For those two selected positions, we used selink software (Cuadros-Espinoza et al. 2022) to compute pairwise F_{ST} between the target population and populations included in the High-SNP density dataset. To avoid sample size bias, we downsampled each population with a sample size larger than 13 individuals. For the downsampled dataset, we repeated XP-EHH-based tests and recalculated allele frequencies for the two top selected SNPs included in one candidate region.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank all the volunteers who participated in this study. In this regard, we are also grateful to Dr. Alioune Dème from the Department of History, Faculty of Letters and Human Sciences, Université Cheikh Anta Diop, Dakar, Senegal, Dr. Warou Abadji Alifa from Centre National d’Appui à la Recherche, N’Djamena, Chad, and Mohammed G. Mokhtar from Arabic Department, Faculty of Arts, University of Kordofan, Al-Ubayyid, Sudan for their invaluable help with the collection of the samples in their countries. We thank Francisco Ceballos, Mario Vicente, and Nina Hollfelder for helpful discussions. The work in this study was funded by grants to V.Č. and C.S. V.Č. was funded by the Grant Agency of the Czech Republic (grant number: 19-09352S-14 P505). C.S. was funded by the European Research Council (ERC StG AfricanNeo, grant no. 759933) and the Knut and Alice Wallenberg Fellowship grant. C.F.-L. was funded in part by the European Research Council (ERC StG AfricanNeo, grant no. 759933 to C.S.), the Sven and Lilly Lawski’s Foundation, the Bertil Lundman’s Foundation, the Marcus Borgström Foundation and the Royal Physiographic Society of Lund (Nilsson-Ehle Endowments). The genotyping was performed by the SNP&SEQ Technology Platform (Uppsala, Sweden). The facility is part of the National Genomics Infrastructure supported by the Swedish Research Council for Infrastructures and Science for Life Laboratory (NGI-SciLifeLab), Sweden. The SNP&SEQ Technology Platform is also supported by the Knut and Alice Wallenberg Foundation. Computations on genome-wide SNP data were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX (project numbers: snic2019/8-230 and snic2021/22-573), partially funded by the Swedish Research Council (through grant agreement no. 2018-05973).

Author Contributions

V.Č. and C.S. conceived, designed, and coordinated the study. V.Č., M.Č., E.P., and M.Y.D. collected the samples in four African countries from the Sahel/Savannah belt. C.F.L., P.T., and M.Y.D. performed bioinformatic analyses. M.Č., E.P., and M.Y.D. performed laboratory analysis. C.F.L., V.Č., and C.S. wrote the article with the input of the other authors.

Data Availability

The genome-wide SNP data generated for the first time in this study will be available for academic research use through the ArrayExpress repository (accession number: E-MTAB-12243). An authorized NIH Data Access Committee (DAC) granted data access to Carina Schlebusch for the controlled-access genetic data analyzed in this study that were previously deposited by Scheinfeldt et al. (2019) in the NIH dbGAP repository (dbGaP accession code: phs001780.v1.p1; project approval date: 2019-05-17).

References

Consortium THR, the Haplotype Reference Consortium. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* **48**(10):1279–1283.

Abu-Amero KK, González AM, Larruga JM, Bosley TM, Cabrera VM. 2007. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. *BMC Evol Biol.* **7**:32.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinf.* **12**:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.***19**(9):1655–1664.

Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, Fadhlaoui-Zid K, Zalloua P, Hellenthal G, Comas D. 2017. Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Mol Biol Evol.* **34**(2):318–329.

Armitage SJ, Bristow CS, Drake NA. 2015. West African monsoon dynamics inferred from abrupt fluctuations of Lake Mega-Chad. *Proc Natl Acad Sci U S A.* **112**(28):8543–8548.

Asad T. 1970. *The Kababish Arabs: power, authority and consent in a nomadic tribe*. Westport, USA: Praeger.

Ashenhurst JR, Harden KP, Mallard TT, Corbin WR, Fromme K. 2017. Developmentally specific associations between CNR1 genotype and Cannabis use across emerging adulthood. *J Stud Alcohol Drugs*. **78**(5):686–695.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, *et al*. 2015. A global reference for human genetic variation. *Nature* **526**:68–74.

Quint. 2006. Do you speak Kordofanian? In: *7th International Sudan studies conference*. Norway: University of Bergen. Available from: https://halshs.archives-ouvertes.fr/halshs-00171745/.

Bakker D, Müller A, Velupillai V, Wichmann S, Brown CH, Brown P, Egorov D, Mailhammer R, Grant A, Holman EW. 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typol*. **13**(1):169–181.

Beins EC, Beiert T, Jenniches I, Hansen JN, Leidmaa E, Schrickel JW, Zimmer A. 2021. Cannabinoid receptor 1 signalling modulates stress susceptibility and microglial responses to chronic social defeat stress. *Transl Psychiatry*. **11**(1):164.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. **57**(1):289–300.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: the impact of rare variants. *Genome Res*. **23**(9):1514–1521.

Brito JC, Martínez-Freiría F, Sierra P, Sillero N, Tarroso P. 2011. Crocodiles in the Sahara desert: an update of distribution, habitats and population status for conservation planning in Mauritania. *PLoS One*. **6**(2):e14734.

Browning BL, Browning SR. 2011. A fast, powerful method for detecting identity by descent. *Am J Hum Genet*. **88**(2):173–182.

Browning SR, Browning BL. 2015. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet*. **97**(3):404–418.

Brunk K, Gronenbom D. 2004. Floods, droughts, and migrations. The effects of Late Holocene lake level oscillations and climate fluctuations on the settlement and political history in the Chad Basin. In: Krings M, Platte E, editors. *Living with the lake. Perspectives on history, culture and economy of Lake Chad*. Cologne, Germany: Rüdiger Köppe Verlag. p. 101–132.

Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, Amenga-Etego LN, Enimil A, Apinjoh T, Ndila CM, *et al*. 2016. Admixture into and within sub-Saharan Africa. *Elife*. **5**: e15266.

Butali A, Mossey PA, Adeyemo WL, Eshete MA, Gowans LJJ, Busch TD, Jain D, Yu W, Huan L, Laurie CA, *et al*. 2019. Genomic analyses in African populations identify novel risk loci for cleft palate. *Hum Mol Genet*. **28**(6):1038–1051.

Carré M, Azzoug M, Zaharias P, Camara A, Cheddadi R, Chevalier M, Fiorillo D, Gaye AT, Janicot S, Khodri M, *et al*. 2019. Modern drought conditions in western Sahel unprecedented in the past 1600 years. *Clim Dyn*. **52**:1949–1964.

Ceballos FC, Hazelhurst S, Ramsay M. 2019. Correction to: runs of homozygosity in sub-Saharan African populations provide insights into complex demographic histories. *Hum Genet*. **138**(10):1143–1144.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet*. **19**(4):220–234.

Černý V, Čížková M, Poloni ES, Al-Meeri A, Mulligan CJ. 2016. Comprehensive view of the population history of Arabia as inferred by mtDNA variation. *Am J Phys Anthropol*. **159**(4): 607–616.

Černý V, Fortes-Lima C, Tříska P. 2021. Demographic history and admixture dynamics in African Sahelian populations. *Hum Mol Genet*. **30**:R29–R36.

Černý V, Kulichová I, Poloni ES, Nunes JM, Pereira L, Mayor A, Sanchez-Mazas A. 2018. Genetic history of the African Sahelian populations. *Hladnikia (Ljubl)*. **91**(3):153–166.

Černý V, Mulligan CJ, Rídl J, Zˇaloudková M, Edens CM, Hájek M, Pereira L. 2008. Regional differences in the distribution of the sub-Saharan, West Eurasian, and South Asian mtDNA lineages in Yemen. *Am J Phys Anthropol*. **136**(2):128–137.

Černý V, Pereira L, Musilová E, Kujanová M, Vašíková A, Blasi P, Garofalo L, Soares P, Diallo I, Brdička R, *et al*. 2011. Genetic structure of pastoral and farmer populations in the African Sahel. *Mol Biol Evol*. **28**(9):2491–2500.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**:7.

Chen J, Bardes EE, Aronow BJ, Jegga AG. 2009. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. **37**:W305–W311.

Chiang CWK, Ralph P, Novembre J. 2016. Conflation of short identity-by-descent segments bias their inferred length distribution. *G3* **6**(5):1287–1296.

Čížková M, Munclinger P, Diallo MY, Kulichová I, Mokhtar MG, Dème A, Pereira L, Černý V. 2017. Genetic structure of the Western and Eastern African Sahel/Savannah Belt and the role of nomadic pastoralists as inferred from the variation of D-loop mitochondrial DNA sequences. *Hum Biol*. **89**(4):281–302.

Clarke T-K, Bloch PJ, Ambrose-Lanci LM, Ferraro TN, Berrettini WH, Kampman KM, Dackis CA, Pettinati HM, O'Brien CP, Oslin DW, *et al*. 2013. Further evidence for association of polymorphisms in the CNR1 gene with cocaine addiction: confirmation in an independent sample and meta-analysis. *Addict Biol*. **18**(4):702–708.

Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016. Model-free estimation of recent genetic relatedness. *Am J Hum Genet*. **98**(1):127–148.

Cuadros-Espinoza S, Laval G, Quintana-Murci L, Patin E. 2022. The genomic signatures of natural selection in admixed human populations. *Am J Hum Genet*. **109**(4):710–726.

Cunnison I. 1966. *Baggara Arabs: power and the lineage in a Sudanese nomad tribe*. Oxford, UK: Clarendon Press.

Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. 2013. Haplotype estimation using sequencing reads. *Am J Hum Genet*. **93**(4): 687–696.

Diallo MY, Čížková M, Kulichová I, Podgorná E, Priehodová E, Nováčková J, Fernandes V, Pereira L, Černý V. 2022. Circum-Saharan prehistory through the Lens of mtDNA diversity. *Genes (Basel)*. **13**(3):533.

Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. 2019. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet*. **15**(11): e1008432.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, *et al*. 2012. The UCSC genome browser database: extensions and updates 2011. *Nucleic Acids Res*. **40**:D918–D923.

Enattah NS, Jensen TGK, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, *et al*. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am J Hum Genet*. **82**(1):57–72.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet*. **30**(2):233–237.

Faraji A, Dehghan Manshadi HR, Mobaraki M, Zare M, Houshmand M. 2016. Association of ABCB1 and SLC22A16 gene polymorphisms with incidence of doxorubicin-induced febrile neutropenia: a survey of Iranian breast cancer patients. *PLoS One*. **11**(12): e0168519.

Feng Q, Lu D, Xu S. 2018. Ancestrypainter: a graphic program for displaying ancestry composition of populations and individuals. *Genomics Proteomics Bioinformatics*. **16**(5):382–385.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol*. **128**(2):415–423.

Fernandes V, Brucato N, Ferreira JC, Pedro N, Cavadas B, Ricaut F-X, Alshamali F, Pereira L. 2019. Genome-wide characterization of

Arabian Peninsula populations: shedding light on the history of a fundamental bridge between continents. *Mol Biol Evol*. **36**(3): 575–586.

François O, Martins H, Caye K, Schoville SD. 2016. Controlling false discoveries in genome scans for selection. *Mol Ecol*. **25**(2): 454–469.

Fortes-Lima C, Gessain A, Ruiz-Linares A, Bortolini M-C, Migot-Nabias F, Bellis G, Moreno-Mayar JV, Restrepo BN, Rojas W, Avendaño-Tamayo E, *et al.* 2017. Genome-wide ancestry and demographic history of African-descendant Maroon communities from French Guiana and Suriname. *Am J Hum Genet*. **101**(5):725–736.

Fuller DQ, Hildebrand E. 2013. Domesticating plants in Africa. In: Mitchell P, Paul, editors. *The Oxford handbook of African archaeology*. Oxford, UK: Oxford University Press.

Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol Ecol Resour*. **17**(1):78–90.

Gautier M, Vitalis R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*. **28**(8):1176–1177.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, *et al.* 2015. The African genome variation project shapes medical genetics in Africa. *Nature* **517**(7534):327–332.

Haber M, Mezzavilla M, Bergström A, Prado-Martinez J, Hallast P, Saif-Ali R, Al-Habori M, Dedoussis G, Zeggini E, Blue-Smith J, *et al.* 2016. Chad genetic diversity reveals an African history marked by multiple Holocene Eurasian migrations. *Am J Hum Genet*. **99**(6):1316–1324.

Hampshire KR, Smith MT. 2001. Consanguineous marriage among the Fulani. *Hum Biol*. **73**(4):597–603.

Hernández CL, Pita G, Cavadas B, López S, Sánchez-Martínez LJ, Dugoujon J-M, Novelletto A, Cuesta P, Pereira L, Calderón R. 2020. Human genomic diversity where the Mediterranean joins the Atlantic. *Mol Biol Evol*. **37**(4):1041–1055.

Hollfelder N, Babiker H, Granehäll L, Schlebusch CM, Jakobsson M. 2021. The genetic variation of lactase persistence alleles in Sudan and South Sudan. *Genome Biol Evol*. **13**(5):evab065.

Hollfelder N, Schlebusch CM, Günther T, Babiker H, Hassan HY, Jakobsson M. 2017. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. *PLoS Genet*. **13**(8):e1006976.

Homewood K. 2008. *Ecology of African pastoralist societies*. Athens, USA: Ohio University Press.

Huang M, Chen Y, Han D, Lei Z, Chu X. 2019. Role of the zinc finger and SCAN domain-containing transcription factors in cancer. *Am J Cancer Res*. **9**(5):816.

Jesse F, Keding B, Lenssen-Erz T, Pöllath N. 2013. I hope your cattle are well': archaeological evidence for early cattle-centred behaviour in the Eastern Sahara of Sudan and Chad. *Pastoralism in Africa: Past, Present and Future*. 66–103.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28**(1):27–30.

Kulichová I, Fernandes V, Deme A, Nováčková J, Stenzl V, Novelletto A, Pereira L, Černý V. 2017. Internal diversification of non-Sub-Saharan haplogroups in Sahelian populations and the spread of pastoralism beyond the Sahara. *Am J Phys Anthropol*. **164**(2):424–434.

Kulichová I, Mouterde M, Mokhtar MG, Diallo I, Tříska P, Diallo YM, Hofmanová Z, Poloni ES, Černý V. 2021. Demographic history was a formative mechanism of the genetic structure for the taste receptor TAS2R16 in human populations inhabiting Africa's Sahel/Savannah Belt. *Am J Phys Anthropol*. **177**(3):540–555.

Kuper R, Kröpelin S. 2006. Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. *Science* **313**(5788): 803–807.

Kuper R, Riemer H. 2013. Herders before pastoralism: prehistoric prelude in the Eastern Sahara. In: Bollig M Schnegg M, Wotzka H-P, editors. *Pastoralism in Africa: past, present and future*. New York, USA: Berghahn Books. p. 31–65.

Lima-Junior J da C, Pratt-Riccio LR. 2016. Major histocompatibility complex and malaria: focus on plasmodium vivax infection. *Front Immunol*. **7**:13.

Linseele V. 2013. From first stock keepers to specialised pastoralists in the West African savannah. In: Bollig M Schnegg M, Wotzka H-P, editors. *Pastoralism in Africa: past, present and future*. New York, USA: Berghahn Books. p. 145–170.

Lipson M. 2020. Applying f-statistics and admixture graphs: theory and examples. *Mol Ecol Resour*. **20**(6):1658–1667.

Liu M, Dong J, Ouyang J, Zhao L, Liang G, Shang H. 2019. Metalloprotease TRABD2A restriction of HIV-1 production in monocyte-derived dendritic cells. *AIDS Res Hum Retroviruses*. **35**(10):887–889.

Liu C-C, Shringarpure S, Lange K, Novembre J. 2020. Exploring population structure with admixture models and principal component analysis. *Methods Mol Biol*. **2090**:67–86.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**(4): 1233–1254.

Magnavita C, Dangbet Z, Bouimon T. 2019. The Lake Chad region as a crossroads: an archaeological and oral historical research project on early Kanem-Borno and its intra-African connections. *Afrique: Archéologie & Arts* **15**:97–110.

Maley J, Vernet R. 2015. Populations and climatic evolution in North Tropical Africa from the end of the neolithic to the Dawn of the modern era. *African Archaeol Rev*. **32**(2):179–232.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**(22):2867–2873.

Manning K, Timpson A. 2014. The demographic response to holocene climate change in the Sahara. *Quat Sci Rev*. **101**:28–35.

Mansour-Hendili L, Aissat A, Badaoui B, Sakka M, Gameiro C, Ortonne V, Wagner-Ballon O, Pissard S, Picard V, Ghazal K, *et al.* 2020. Exome sequencing for diagnosis of congenital hemolytic anemia. *Orphanet J Rare Dis*. **15**(1):180.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res*. **27**(2):209–220.

Marcus J, Ha W, Barber RF, Novembre J. 2021. Fast and flexible estimation of effective migration surfaces. *Elife* **10**:e61927.

McIntosh SK. 2020. Long-distance exchange and urban trajectories in the first millennium AD: case studies from the middle Niger and middle Senegal river valleys. *In: Sterry M, Mattingly D, editors. Urbanisation state form ancient Sahara beyond*. Cambridge, UK: Cambridge University Press. p. 521-563.

McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko T, Corre T, Davies G, Kaakinen M, Lyytikäinen L-P, *et al.* 2012. Evidence of inbreeding depression on human height. *PLoS Genet*. **8**(7):e1002655.

Mulder N, Abimiku A, Adebamowo SN, de Vries J, Matimba A, Olowoyo P, Ramsay M, Skelton M, Stein DJ. 2018. H3Africa: current perspectives. *Pharmgenomics Pers Med*. **11**:59–66.

Murphy ACH, Young PW. 2015. The actinin family of actin crosslinking proteins - a genetic perspective. *Cell Biosci*. **5**:49.

Nováčková J, Čížková M, Mokhtar MG, Duda P, Stenzl V, Tříska P, Hofmanová Z, Černý V. 2020. Subsistence strategy was the main factor driving population differentiation in the bidirectional corridor of the African Sahel. *Am J Phys Anthropol*. **171**(3):496–508.

Novembre J, Williams R, Pourreza H, Wang Y, Carbonetto P. 2019. PCAviz: Visualizing principal components analysis. Available from: http://github.com/NovembreLab/PCAviz.

Nychka D, Furrer R, Paige J, Sain S. 2005. *Fields: tools for spatial data*. National Center for Atmospheric Research.

Ouyang X, Becker Jr E, Bone NB, Johnson MS, Craver J, Zong WX, Darley-Usmar VM, Zmijewski JW, Zhang J. 2021. ZKSCAN3 in severe bacterial lung infection and sepsis-induced immunosuppression. *Lab Invest*. **101**(11):1467–1474.

Pabalan N, Chaweeborisuit P, Tharabenjasin P, Tasanarong A, Jarjanazi H, Eiamsitrakoon T, Tapanadechopone P. 2021. Associations of CB1 cannabinoid receptor (CNR1) gene polymorphisms with risk for alcohol dependence: evidence from meta-analyses of genetic and genome-wide association studies. *Medicine (Baltimore)*. **100**(43):e27343.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* **192**(3):1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. **2**(12):e190.

Pedersen J, Benjaminsen TA. 2008. One leg or two? Food security and pastoralism in the Northern Sahel. *Hum Ecol*. **36**(1):43–57.

Pereira L, Černý V, Cerezo M, Silva NM, Hájek M, Vasíková A, Kujanová M, Brdicka R, Salas A. 2010. Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African sahel. *Eur J Hum Genet*. **18**:915–923.

Peter BM, Petkova D, Novembre J. 2020. Genetic landscapes reveal how human genetic diversity aligns with geography. *Mol Biol Evol*. **37**(4):943–951.

Petkova D, Novembre J, Stephens M. 2016. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet*. **48**(1):94–100.

Phelps LN, Broennimann O, Manning K, Timpson A, Jousse H, Mariethoz G, Fordham DA, Shanahan TM, Davis BAS, Guisan A. 2020. Reconstructing the climatic niche breadth of land use for animal production during the African Holocene. *Glob Ecol Biogeogr*. **29**(1):127–147.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. Ancient west Eurasian ancestry in southern and Eastern Africa. *Proc Natl Acad Sci U S A*. **111**(7): 2632–2637.

Podgorná E, Diallo I, Vangenot C, Sanchez-Mazas A, Sabbagh A, Černý V, Poloni ES. 2015. Variation in NAT2 acetylation phenotypes is associated with differences in food-producing subsistence modes and ecoregions in Africa. *BMC Evol Biol*. **15**:263.

Priehodová E, Abdelsawy A, Heyer E, Černý V. 2014. Lactase persistence variants in Arabia and in the African Arabs. *Hum Biol*. **86**(1):7–18.

Priehodová E, Austerlitz F, Čížková M, Mokhtar MG, Poloni ES, Černý V. 2017. The historical spread of Arabian pastoralists to the eastern African Sahel evidenced by the lactase persistence- 13,915* G allele and mitochondrial DNA. *Am J Hum Biol*. **29**(3):e22950.

Priehodová E, Austerlitz F, Čížková M, Nováčková J, Ricaut F-X, Hofmanová Z, Schlebusch CM, Černý V. 2020. Sahelian pastoralism from the perspective of variants associated with lactase persistence. *Am J Phys Anthropol*. **173**(3):423–436.

Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, *et al.* 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet*. **94**(4):496–510.

Rogowski K, Van Dijk J, Magiera MM, Bosc C, Deloulme JC, Bosson A, Peris L, Gold ND, Lacroix B, Grau MB, *et al.* 2010. A family of protein-deglutamylating enzymes associated with neurodegeneration. *Cell* **143**(4):564–578.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164):913–918.

Sanchez-Mazas A, Černý V, Di D, Buhler S, Podgorná E, Chevallier E, Brunet L, Weber S, Kervaire B, Testi M, *et al.* 2017. The HLA-B landscape of Africa: signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection. *Mol Ecol*. **26**(22):6238–6252.

Scheinfeldt LB, Soi S, Lambert C, Ko W-Y, Coulibaly A, Ranciaro A, Thompson S, Hirbo J, Beggs W, Ibrahim M, *et al.* 2019. Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. *Proc Natl Acad Sci U S A*. **116**(10):4166–4175.

Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol*. **5**(7):e171.

Tournebize R, Chu G, Moorjani P. 2022. Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals. *PLoS Genet*. **18**(6):e1010243.

Triska P, Soares P, Patin E, Fernandes V, Černý V, Pereira L. 2015. Extensive admixture and selective pressure across the Sahel Belt. *Genome Biol Evol*. **7**(12):3484–3495.

Turner MD, Schlecht E. 2019. Livestock mobility in sub-Saharan Africa: a critical review. *Pastoralism* **9**(1):13.

Vicente M, Priehodová E, Diallo I, Podgorná E, Poloni ES, Černý V, Schlebusch CM. 2019. Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. *BMC Genomics* **20**(1):915.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. **4**(3):e72.

Wichmann S, Holman EW, Stadler BCH. 2020. The ASJP Database (version 19). Available from: https://asjp.clld.org.

Winchell F, Stevens CJ, Murphy C, Champion L, Fuller D. 2017. Evidence for Sorghum domestication in Fourth Millennium BC Eastern Sudan: spikelet morphology from ceramic impressions of the Butana Group. *Curr Anthropol*. **58**(5):673–683.

Young WC. 1996. In: Spindler LS, Spindler GD, editors. *The Rashaayda Bedouin: Arab Pastoralists of Eastern Sudan*. Fort Worth, Texas: Harcourt Brace College Professors.

Yu Y, Ouyang Y, Yao W. 2018. Shinycircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **34**(7): 1229–1231.