Patterns of population structure and genetic variation within the Saudi Arabian population

D.K. Malomane[1*], M. P. Williams[2], C.D. Huber[2], S. Mangul[3,‡], M. Abedalthagafi[4,5,‡], C. W. K. Chiang[1,6,‡**]

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA.

[2]Department of Biology, Pennsylvania State University, University Park, PA.

[3]Titus Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, University of Southern California, Los Angeles, CA.

[4]Department of Pathology and Laboratory Medicine, Emory University Hospital, Atlanta, GA.

[5]Genomics Research Department, King Fahad Medical City, Riyadh, Saudi Arabia.

[6]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA.

[‡]These authors contributed equally to supervise this work.

[*]Correspondence: kholo642@gmail.com.

[**]Correspondence: charleston.chiang@med.usc.edu.

ABSTRACT

The Arabian Peninsula is considered the initial site of historic human migration out of Africa. The modern-day indigenous Arabians are believed to be the descendants who remained from the ancient split of the migrants into Eurasia. Here, we investigated how the population history and cultural practices such as endogamy have shaped the genetic variation of the Saudi Arabians. We genotyped 3,352 individuals and identified twelve genetic sub-clusters that corresponded to the geographical distribution of different tribal regions, differentiated by distinct components of ancestry based on comparisons to modern and ancient DNA references. These sub-clusters also showed variation across ranges of the genome covered in runs of homozygosity, as well as differences in population size changes over time. Using 25,488,981 variants found in whole genome sequencing data (WGS) from 302 individuals, we found that the Saudi tend to show proportionally more deleterious alleles than neutral alleles when compared

to Africans/African Americans from gnomAD (e.g. a 13% increase of deleterious alleles annotated by AlphaMissense between 0.5 - 5% frequency in Saudi, compared to 7% decrease of the benign alleles; $P < 0.001$). Saudi sub-clusters with greater inbreeding and lower effective population sizes showed greater enrichment of deleterious alleles as well. Additionally, we found that approximately 10% of the variants discovered in our WGS data are not observed in gnomAD; these variants are also enriched with deleterious annotations. To accelerate studying the population-enriched deleterious alleles and their health consequences in this population, we made available the allele frequency estimates of 25,488,981 variants discovered in our samples. Taken together, our results suggest that Saudi's population history impacts its pattern of genetic variation with potential consequences to the population health. It further highlights the need to sequence diverse and unique populations so to provide a foundation on which to interpret medical- and pharmaco- genomic findings from these populations.

INTRODUCTION

Saudi Arabia is the largest country in the Arabian Peninsula (AP), the central hub of the world that connects Africa, Asia and Europe. The AP is considered one of the initial sites of historic human migration out of Africa (OOA), with presence of human footprints reported at least since 50 – 60 thousand years ago (kya) and as early as 85 – 120 kya [1–6]. The contribution of the earliest expansion in present-day Arabians or other modern non-Africans has not been fully explored. However, genetic evidence suggests that all present-day Middle Eastern populations predominately descend from the same ancestral OOA population, as is the case for the other non-Africans [6].

The genetic diversity of today's Arabians is shaped by a complexity of ancestries from historic and recent splits and admixture events. An early divergence of Arabian ancestors from other non-Africans is estimated to have happened shortly after the OOA event [3]. Among the non-Africans, Arabians carry a higher proportion of a deeply diverged 'ghost' ancestry, labeled 'Basal Eurasian,' and lower levels of Neanderthal admixture [7,8]. It has been hypothesized that the Arabians descended from the Basal Eurasians which diverged from other non-Africans before the major Neanderthal admixture [7–10]. Alternatively, the Basal Eurasians diverged from the non-Africans shortly after the OOA and was isolated until experiencing a later admixture in the Middle East around 25kya, which diluted the Neanderthal ancestry [11]. Since the OOA, Arabians have experienced series of admixtures, and the present-day Arabians have shared ancestries with various groups including Africans, South Asians, Europeans, Levantines, and Iranians [6,12,13].

Despite the rich history of ancestries and being in the center of the world, for centuries the genetic pool of the Arab countries and the Greater Middle East (GME) have been greatly influenced and refined by mating practices. Arab countries have a high rate of endogamous and consanguineous marriages [14,15], especially in Saudi Arabia with rates as high as 58% [16,17]. These endogamous marriages are meant to preserve family structure and strengthen bonds, as well as to ensure cultural, religious, financial and social stability [17–19]. Many of the consanguineous marriages are found among close relatives (e.g. 28.4% among first cousins[16]), but can also

extend to members of the same or related tribal groups. Endogamy leads to regional genetic isolation and population substructure. A recent study analyzing the population structure of Saudi Arabia based on less than a thousand indigenous genotyped samples showed a signature of tribal stratification within the population [20]. Furthermore, because many deleterious mutations are recessive-acting, consanguineous unions have the potential of increasing the burden of deleterious alleles in a population as these deleterious recessive alleles are co-inherited in offsprings [14,21]. This could increase the prevalence of genetic disorders, some of which have indeed been observed in Saudi Arabia [22,23]. While in the long run these deleterious recessive alleles are likely exposed to purifying selection due to increased homozygosity [24,25], previous studies in the GME region have found no evidence of genetic purging of deleterious alleles due to the long-term practices of endogamy and consanguinity. Instead, intense inbreeding and/or reproductive compensation have been suggested to counteract the effectiveness of purifying selection in consanguineous populations [26–29]. Moreover, with small effective population sizes and inbreeding, variants acting additively tend to accumulate at a much higher rate and negative selection is less effective in removing weakly deleterious alleles [30–33]. Overall, small effective population size and intense inbreeding through consanguinity may result in an abundance of deleterious alleles due to its negative impact on the effectiveness of negative selection.

In the present study, we genotyped 3,352 individuals with high-density SNP array and whole genome sequenced (WGS) 302 individuals to investigate how the population history and cultural practices have shaped the genetic structure of the Saudi population. We investigated the pattern of admixture in Saudi sub-populations through the lens of both modern and available ancient DNA samples and inferred the population size trajectories over time. Finally, we leveraged the 302 whole genome sequenced individuals to further explore the impact of the population history on the distribution of genetic variation within social structure and potential consequences to today's population health.

RESULTS

Genetic substructure and admixture patterns of Saudi Arabians

We merged 3,352 genotyped Saudi individuals after quality control (see **Methods**) with 302 whole genome sequencing (WGS) samples, based on 603,833 shared segregating sites, to explore the population structure. We performed principal component analysis (PCA) on the combined set and projected the first 10 principal components (PCs; **Figure S1**) down to 2 dimensions using Uniform Manifold Approximation and Projection (UMAP). Average Silhouette Width (ASW) clustering on the UMAP results suggested that twelve genetic sub-clusters within the Saudi population best fit the data, although visually 6 to 8 sub-clusters may also be sensible (**Methods; Figure 1A**). The distribution of individuals by ASW clusters are presented in **Figure S2A.** To aid in the geographical interpretation of these sub-populations, we intersected the clustering results with self-reported or predicted tribal geographic labels from the cohort (**Methods**). Due to privacy protection and ethical restrictions, we did not have access to specific tribal name of each individual but rather the geographic regions of the tribes. We found that the 12 clusters corresponded to geographical structure of the tribes within Saudi Arabia, with each cluster generally consisting of a majority of its members from a single geographical region (Central, West, North, South, or East) whether using harmonized tribal labels or self-reported labels when available, except for clusters 11 and 12 (**Table S1, Figure 1A** and **Figure S2B**). Both clusters 11 and 12 had multiple dominating tribal regions. We note that there were multiple separate genetic clusters affiliated to the same geographic regions (e.g. clusters 2, 3, and 9 from Central region; 4, 5, 7, and 10 from the Western region, etc.). This observation is unlikely due to errors in inferring tribal regional labels, since previous studies using completely self-identified indigenous tribal information also showed limited inter-tribal marriages within a region [20]. Among the clusters we inferred, cluster5 from the Western region appeared to be most differentiated from the rest of the cohort, in both UMAP (**Figure 1A**) and PCA (PCs 6 and 7; **Figure S1**).

For a global comparison, we compared the Saudi clusters to the populations from the Human Genome Diversity Panel (HGDP) [34]. Consistent with previous reports [20,35], the Saudi individuals clustered between Africans, Central & South Asians and Europeans and were the most distant to East Asians (**Figure 1B**). Cluster12 showed the strongest affinity towards the African reference individuals, followed by cluster3, while cluster11 showed affinity towards Europeans, Africans and Central & South Asians. The remaining clusters co-localized mainly with the Middle Eastern reference individuals from the HGDP panel.

Our observation of population structure from PCA is also supported by unsupervised ADMIXTURE analysis combining Saudi with HGDP populations. For instance, at K = 4, clusters 11 and 12 also exhibited the highest levels of admixture (**Figure 1C** and **Table S2**). We labeled ancestry components by the HGDP population with dominating or highest admixture proportions (**Table S2**). The most dominating ancestry in the Saudi clusters was one largely shared with the HGDP Middle Eastern populations (Druze, Bedouin, Mozabite, and Palestinian (**Figure 1C**, top), which we termed Middle Eastern-like (ME-like; cyan) ancestry, with Bedouin showing the largest amount of this ancestry among the HGDP Middle Eastern individuals (**Table S2**). On the other hand, clusters 12, 11, and 3 had on average less than two-thirds of this ME-like ancestry component and were enriched with African (AFR)- (red) and/or European (EUR)-like (green) ancestries.

One of the lowest cross validation errors occurred at K = 9 (**Figure S3C**), which also introduced a new ancestry component distinguishing the CSA-like ancestry from the EUR-like ancestry (For K = 5 to 8 (**Figure S3A**) the ADMIXTURE algorithm were mostly distinguishing ancestries within Saudi Arabian clusters themselves). At K = 9 (**Figure 1C** bottom, and **Table S3**), the relationship between cluster11 and the Central & South Asians (CSA) that we observed on PCA can also be observed, where cluster11 carried more (average proportion = 0.223) of such CSA-like ancestry compared to other clusters (average proportions less than 0.1). We also observed three ME-like ancestries (**Table S3** and **Figure 1C,** bottom). One of the Middle Eastern-like ancestry (ME-2) that is dominant in several (> 10% in 10 out of the 12) clusters, particularly in clusters 6, 8, 1, and 7, is also found in Sardinians and other Italians & Adygei but completely missing in the

Russians. The other Middle Eastern-like ancestries (ME-1 and ME-3) are found distributed in a subset of the clusters. In particular, the ME-3 ancestry is found in high proportions in clusters 10, 4, 5, and 2 (average proportions = 0.45 - 0.92), and was also found in HGDP-Bedouin (but absent from HGDP-Druze, HGDP-Mozabite and HGDP-Palestinian). This ancestry appears to be enriched in Qataris Bedouins and Saudi Arabians but not other Middle Eastern populations, and was suggested to reflect an indigenous Arab ancestry [3].

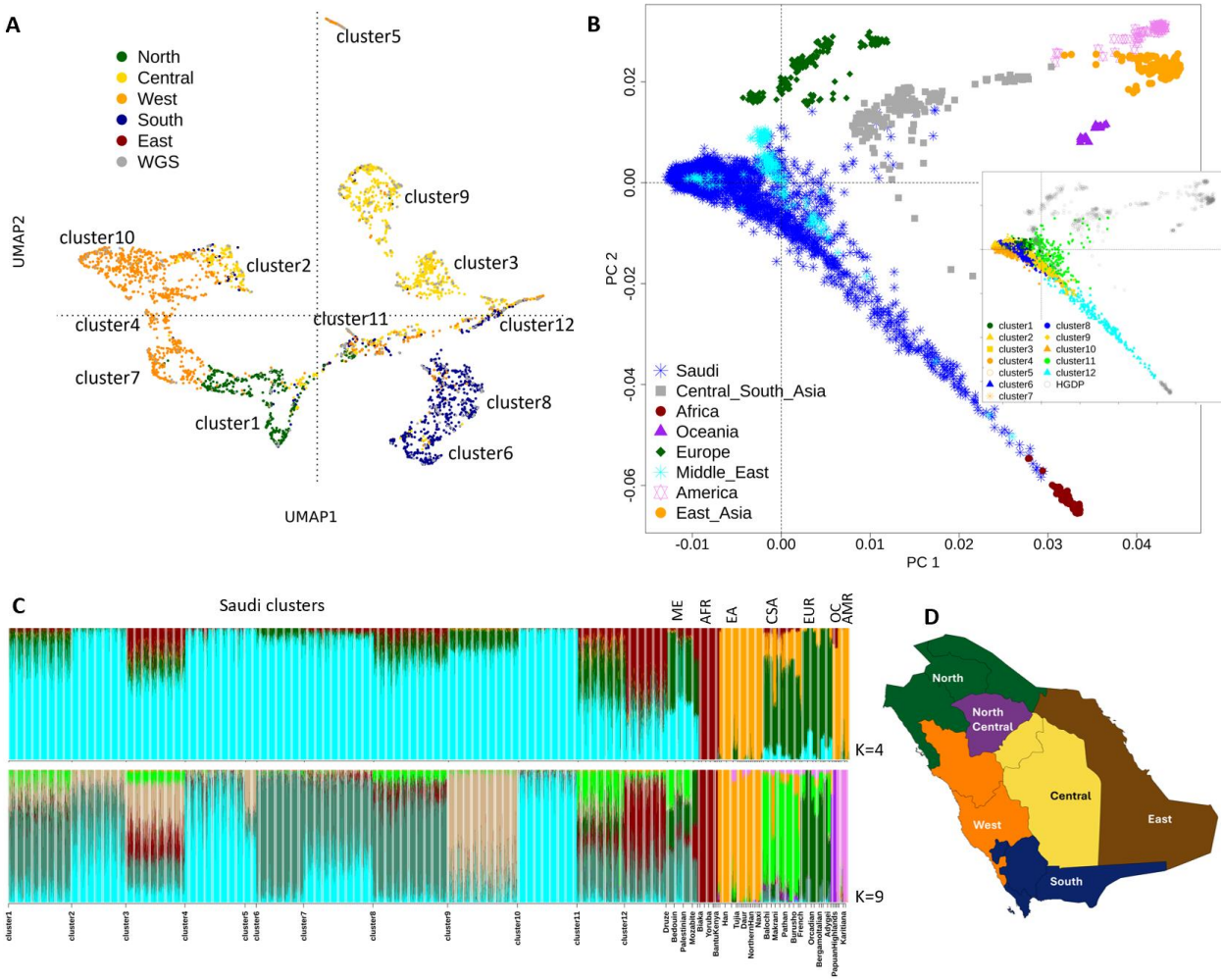


Figure 1. The genetic structure of Saudi Arabians and its relation to global populations. (A) A two-dimensional UMAP of Saudi Arabians based on the top 10 principal components. Each individual is colored based on the affiliated tribal region (see (D)). WGS samples did not have self-reported or harmonized tribal affiliation and are assigned their own color. (B) PCA of Saudi Arabian clusters and HGDP populations. Saudi Arabians are grouped in a single group. Inset

shows clusters 1 - 10 colored according to the most prevalent tribal region represented in the cluster (see **Table S1**). Because clusters 11-12 has no single dominating tribal region, they were assigned distinct separate colors. (C) Admixture analysis of Saudi Arabian clusters and HGDP populations for K = 4 (top) and K = 9 (bottom). ME – Middle Eastern, AFR – African, EA – East Asian, CSA – Central & South Asian, EUR – European, OC – Oceania, AMR – American. The names of Saudi clusters and HGDP populations are shown on the bottom X-axis. However, due to limited space some of the labels for smaller populations from HGDP are omitted. Grouped regional labels are shown on the top X-axis of plots. We show the admixture results of the Saudi clusters alone in **Figure S3B.** (D) A regional map of Saudi Arabia with matching colors to the regional labels in (A) and (B).

The evidence for admixture for clusters 12, 11, and 3, together with clusters 8 and 1 were also corroborated by admixture f_3-statistics. Using all possible pairs of HGDP populations as potential proxies of ancestral sources (possibly through shared ancestry), only these five Saudi sub-clusters showed any significantly negative f_3-statistics indicative of admixture (**Tables S4 - 8**). We further investigated the degree of shared drift between the Saudi clusters and the HGDP populations using the outgroup f_3-statistics. We used HGDP-Han population as the outgroup over the typical choices of the San, Mbuti, or Yoruba populations since we expect African being a plausible admixing source in Saudi and indeed found positive gene flow between the HGDP African populations and the Saudi clusters (**Methods; Figure 1C**). We found the pattern of the outgroup f_3-statistics to be similar among Saudi clusters 1 - 10, in contrast to clusters 11 and 12 (**Figure S4**). Clusters 1 - 10 showed highest shared drift to other Saudi clusters followed by the Middle Eastern and European populations (**Figure S4**). On the other hand, cluster12 showed most shared drift to HGDP-African populations, even more so than it is to other Saudi clusters, further corroborating the results in **Figure 1B** and **C**. In relation to other Middle Eastern populations from HGDP, clusters 1 - 10 showed greater shared drift to Bedouin while cluster12 was most related to the Mozabite. Given the high African ancestry in cluster12, this relationship with Mozabite population from North Africa is not surprising, and could have either arose from shared sub-Saharan ancestry or an Arabic admixture event during the Islamic expansion into North Africa about 1,200 – 1,400 kya [36], or a combination of both. Interestingly, in relation to the European populations, all Saudi clusters shared the greatest drift with the Sardinians, and

the least with the Russians. The relationship to the Sardinians is in concordance with Charati et al., [37]. Sardinians heavily harbor early farmer Neolithic ancestry, which expanded into Europe from the Near East and Anatolia [38–40]. Saudi Arabians are estimated to have split from Sardinians around 20 kya [6]. Their relationship might be reflecting the ancient Neolithic ancestry, a product of the Arabian migration into Italian islands or a result of continuous and recent admixtures [37,40–42].

Genetic legacy of ancient ancestry in modern day Saudi Arabians

We expanded our understanding of Saudi genetic history by integrating 302 WGS Saudi individuals with ancient DNA (aDNA) datasets (1240k AADR v54.1 [43] and ancient Bahrain individuals [12]; **Figure S5**; **Methods**). Our integrated analysis with aDNA data corroborated many of our findings above on population structure and admixture history using only modern DNA. Previous studies have shown that Eastern Arabian Peninsula (AP) populations have higher ancient Zagros mountains/Caucasus mountains hunter-gatherer (CHG)-related ancestry than Western AP populations [7,11,44]. We observed a similar geographical divide among Saudi Arabians. Using an Epipaleolithic sample from the Natufian culture (Natufian EpiP) to represent Levantine ancestries and a sample from the Ganj Dareh Neolithic settlement in western Iran (Ganj Dareh N) along with other CHG to represent Zagros ancestries, we found that Saudi clusters mainly from the West and South region of Saudi Arabia (clusters 6, 10, 5, 8, 7, and 4) showed significant (f_4 Z-score < -3) excess shared drift with Levantine ancestries relative to Zagros ancestries (**Figure S6,** bottom, and **Table S9**). We also evaluated the spatiotemporal distribution of African ancestry amongst the Saudi clusters using the outgroup f_3-statistic of form f_3(Han.DG; ancient and present-day African population, Saudi cluster) along with 108 African populations spanning from the present-day to ~15 kya. Temporal analysis across four time bins ([0,0], (0, 1,000 kya], (1,000 kya, 4,000 kya], and (4,000 kya, 15,500 kya]) revealed that Saudi clusters 12 and 3 consistently exhibited elevated African ancestry (**Figure S7 and Table S10**), consistent with **Figures 1B**, **1C**, and **S4**. Importantly, this increased genetic affinity

appears associated with African populations south of approximately 20°N latitude, whereas genetic affinity to Northern African ancestry remains relatively uniform across all time bins and Saudi clusters. This finding is also corroborated by f_4-statistic confirming that Saudi cluster12 shows the highest level of African ancestry (**Figure S6**, top and **Table S11**).

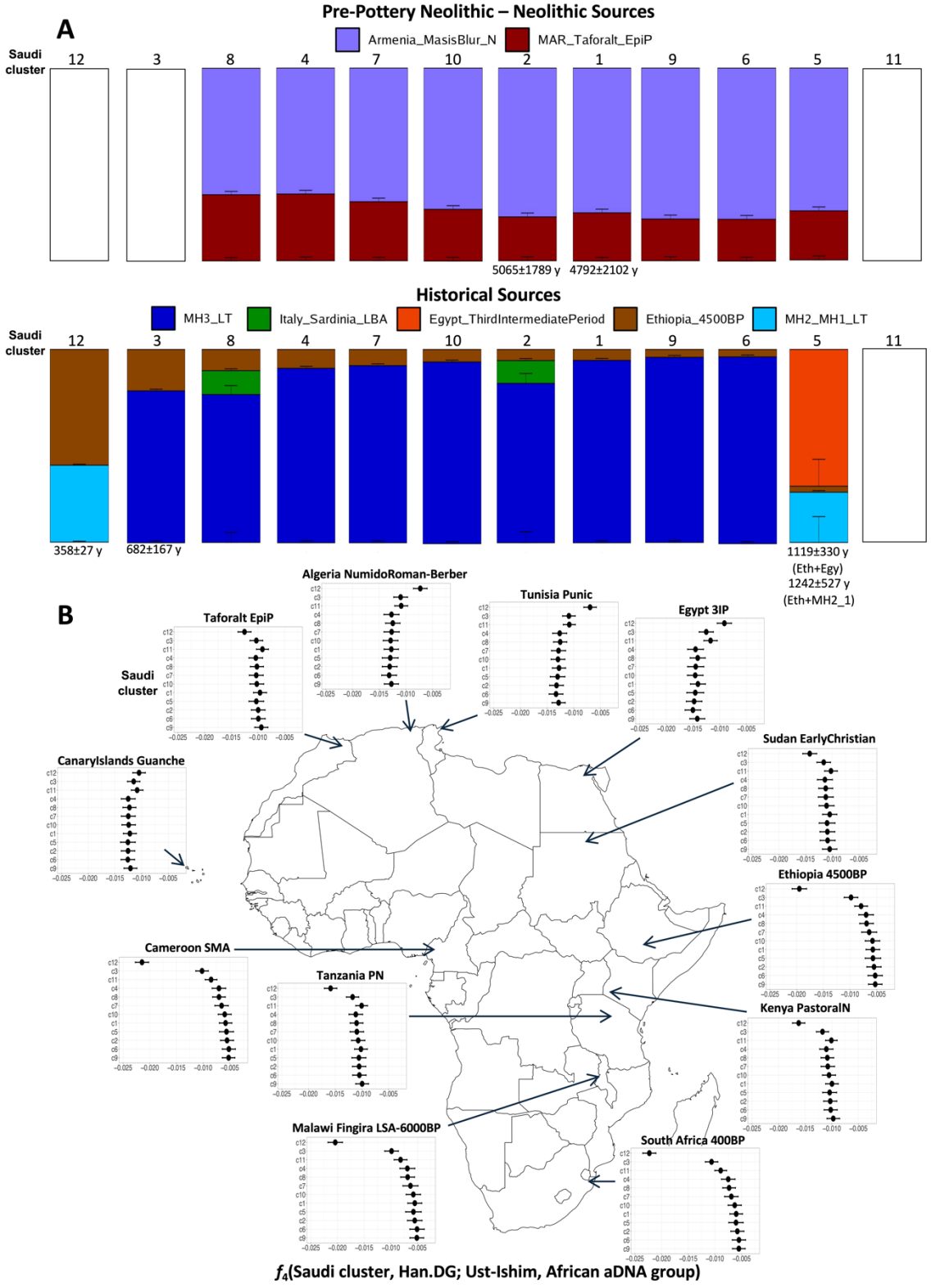
Previous research has identified genetic continuums amongst present-day AP populations with respect to 'Basal Eurasian' ancestry—a hypothesized ghost lineage that diverged from the primary out-of-Africa lineage prior to Neanderthal introgression [10,45]. We used the f_4-statistic of form f_4(Saudi cluster, Han.DG; Ust-Ishim, ancient African population) to estimate the relative amount of Basal Eurasian ancestry (i.e. the drift basal to the shared drift between Han.DG and Ust-Ishim, a 45 ky sample from western Siberia) across the Saudi cluster cohort. Strongly negative f_4 values of this form is consistent with elevated shifted drift with Basal Eurasian ancestry. However, recent African admixture could confound this signal, and different African aDNA samples are likely to represent Basal Eurasian ancestry to different degrees, thereby prompting us to test a range of ancient African populations. We found that most Saudi clusters, regardless of geographical locations, showed similar levels of Basal Eurasian ancestry (**Figure 2B** and **Table S12**). We found that Saudi cluster12, and to a lesser extent cluster3, indeed showed stronger negative f_4 values when the African aDNA source is from Central, Southern, or Eastern parts of Africa, consistent again with their recent African admixture from these regions (**Figure 2B** and **Table S12**). The estimates of Basal Eurasian ancestries across Saudi clusters are much more comparable to each other when using the North African references (**Figure 2B** and **Table S12**), consistent with the recent proposal that the Epipaleolithic Iberomaurusian Taforalt population is the best proxy for Basal Eurasian ancestry [7]. Moreover, these results confirm that the North African aDNA references are not closely related to the admixing African source in Saudi clusters 12, and 3.

To investigate the recourse of Saudi cluster ancestries through time, we implemented qpAdm modeling with chronologically stratified sources across four temporal bins: Paleolithic-Neolithic (P-N), Chalcolithic (C), Bronze Age (BA), and Historical (H), maintaining a fixed set of 11 right groups throughout (**Methods** and **Tables S13 - S16**). Saudi clusters were most successfully

modeled using sources from the temporal bookends, the earliest (P-N) and most recent (H) periods (**Figure 2A**). The poor performance of Chalcolithic and Bronze Age sources in modeling Saudi cluster ancestry (**Table S14 - S15**) likely reflects the current absence of well-represented sources, such as an absence of ancient Arabian Peninsula genetic data from these intermediate periods. In the P-N period a two-source model combining similar proportions of North African (Taforalt EpiP) and Neolithic Armenian (MasisBlur N) components appeared to be the most plausible model for most of the Saudi clusters, corroborating f_4-statistic evidence of approximately equal ancient North African genetic affinity (**Figure 2B**). Clusters 3, 11, and 12 were rejected in this relatively simple model, reflecting their complicated admixture history as described above. Re-analysis of clusters 3, 11, and 12 by removing Mbuti.DG from the qpAdm construct (owing to their greater African-related ancestry) did not result in a better fitted model either (data not shown). We used DATES [46] to estimate admixture timing and only Saudi clusters 1 (4792 ± 2102 years) and 4 (5065 ± 1789 years) returned well-fitting admixture timing estimates (normalized root mean standard deviation (NRMSD) < 0.7, Z-score > 2 and **Table S17**).

Analysis of Saudi cluster ancestry through Historical sources revealed consistent ancestral contributions from two Bahraini groups (MH3 LT and MH1-MH2 LT; MH3 LT has greater Levantine-related ancestry [12]). The MH3 LT + Ethiopia 4500BP model was plausible for seven Saudi clusters: clusters 1, 3, 4, 6, 7, and 9 with only Saudi cluster3 returning a well-fitting admixture date estimate (682 ± 167 years). For Saudi clusters 2 and 8, the inclusion of Sardinia LBA as a third source on top of Ethiopia 4500BP + MH3 LT appeared to improve the model. However, admixture timing analysis revealed high-uncertainty estimates for both two-source and three-source qpAdm models, whereby MH3 LT + Sardinia LBA was well-fitting for only cluster8 (612 ± 835 years) and the Ethiopia 4500BP + Sardinia LBA model well-fitting for Saudi clusters 2 (474 ± 625 years) and 8 (820 ± 952 years). Interestingly the Arabian ancestry component for both Saudi clusters 12 and 5 is best modeled by the MH1-MH2 LT group, characterized by reduced Levantine affinity [12]. Consistent with their high African ancestry, Saudi cluster12 is plausibly modeled possessing 0.60 ± 0.005 Ethiopia 4500BP ancestry, with the estimated timing of their Ethiopia 4500BP + MH1-MH2 LT admixture model at 358 ± 27 years. Ancestry modeling of Saudi cluster5 required a third source component from Egypt's Third

Intermediate Period (Egypt 3IP: 0.71 ± 0.14; MH1-MH2 LT: 0.26 ± 0.13). DATES analysis of Saudi cluster5's ancestry formation revealed well-fitting estimates for Ethiopia_4500BP + Egypt 3IP (1120 ± 330 years) and Ethiopia_4500BP + MH2_MH1_LT (1242 ± 528 years) models, while Egypt + MH2_MH1_LT modeling failed statistical fitting criteria. Finally, the unique ancestry configuration of Saudi cluster11 demonstrated above (**Figures 1C, S3, and S4)**, is also manifest in qpAdm modeling whereby all models across all time periods fit the data poorly. Taken together, these findings suggest the present-day Saudi Arabian ancestry component was formed through multiple ancient non-local ancestry contributions to a predominant local Arabian background – represented by Bahraini groups MH3 LT and MH1-MH2 LT. The majority of clusters showed compatibility with MH3 LT and Ethiopian ancestry, while specific clusters exhibited unique patterns: clusters 2 and 8 incorporated Sardinian ancestry, clusters 12 and 5 showed strong African components with MH1-MH2 LT base, and cluster11 displayed a distinctive genetic profile unable to be plausibly modeled with the current sources, further revealing the importance of future ancient DNA research in this region.



294

Figure 2. Ancestry compositions in Saudi Arabians estimated with aDNA data as reference. (A) Barplots for plausible (p-value ≥ 0.01 and admixture weights between 0 and 1) qpAdm models grouped by age brackets of source populations (top and bottom; **Methods**). For Pre-Pottery Neolithic – Neolithic sources (top), three clusters were rejected under the Armenia_MasisBlur N + MAR_Taforalt_EpiP qpAdm model at the statistical threshold cut-off: cluster12, 3, and 11. We display under the corresponding qpAdm barplot well-fitting (nrmsd < 0.7 and Z > 2) estimates of admixture timing in years. (B) 'Basal Eurasian' ancestry estimated from f_4-statistic of form f_4(Saudi cluster, Han.DG; Ust-Ishim, African aDNA group) with varying ancient African groups [47–55]. We plotted three standard errors for each f_4-statistic. The Saudi cluster (y axis) order in each plot is retained throughout (c12, c3, c11, c4, c8, c7, c10, c1, c5, c2, c6, and c9) following decreasing value for the statistic f_4(Saudi cluster, Han.DG; Ust-Ishim, Ethiopia 4500BP). Significant (absolute Z-score > 3) negative f_4-statistic values indicate the Saudi cluster possesses excess shared drift basal to the shared drift between the groups (Han.DG and Ust-Ishim), commonly interpreted as deriving from a population basal to the OOA event (i.e. the Basal Eurasian).

Genetic variation within Saudi is shaped by the social structure

The sub-clusters of Saudi Arabians identified in this study exhibits a diverse distribution of runs of homozygosity (ROH) between the individuals. The general pattern of number of ROH (NROH) vs. the sum total length of ROH (SROH) showed a distinct relationship with the proportion of ME-like ancestries (**Figure 3A**). Across Saudi clusters, the median total sum of runs of homozygosity ranged from 38.12 Mb to 232.6 Mb, while the median number of ROH ranged from 42 to 150 ROHs. In terms of the mean, cluster12 had the shortest mean length and smallest mean number of ROHs, followed by cluster11 and cluster3 (**Figure 3A**), which is characteristic of larger effective population size and consistent with greater admixture from more diverse African ancestral populations (**Figure 1C**). Cluster5 had the highest burden of ROH, with highest average number and total length of ROH, followed by cluster10, reflecting a consequence of both long-term small effective size and/or consanguinity [56].

We also followed a previous approach [57] and divided the ROHs into three classes based on length: < 635 kb for short ROHs, between 635kb and 1671kb for intermediate ROHs, and > 1671kb for long ROHs (**Methods**). Short ROHs indicate homozygosity from ancient or distant

ancestry, i.e. background relatedness. Intermediate ROHs likely arise from background relatedness with moderate level of inbreeding from past few generations, often due to reduced population sizes or reproductive isolation (e.g. due to geographic or cultural preferences), or from recent bottlenecks followed by recovery. Long ROHs indicate recent inbreeding and are common in populations with high levels of consanguinity [56–58]. When classified by the sizes, we can observe that the overall pattern of NROH vs. SROH (**Figure 3B**) are driven by the long ROHs (**Figure S8**). For both the short and intermediate ROHs, there are clear linear relationships between NROH and SROH (**Figure S8**). In contrast, for the long ROHs, as SROH increase per individual genome, the NROHs are not increasing at the similar linear pattern as observed for short and intermediate ROHs. That is, for individuals with greater SROH due to the long ROHs, they do not have proportionally greater NROH compared to those with less SROH, suggesting that the contributions of SROHs are driven by fewer but longer ROHs in this length class due to recent consanguinity. Therefore, the consequences of consanguinity in not only increasing SROH but also increasing the variance of SROH in a population [56,59]. We also observed the impact of this when considering each Saudi sub-clusters. In general, clusters 12, 11, and 3 have the fewest NROH and the smallest SROH across the ROH classes while clusters 5, 2, and 10 tended to have the most NROH and longest SROHs (**Figure S9A** and **S9B**). The ranked order by both NROH and SROH across the 12 Saudi clusters were very similar for both the short and intermediate length ROHs (**Figure S9A** and **S9B**), but varied for the long ROH class, implying a different pattern of recent inbreeding that differed from ancient demographic events.

To further support the relationship between ROH and ancestry components, we modeled the ROH by ancestry proportion, based on admixture analysis at K = 4 with the HGDP populations. We found that SROH increase with the increase in ME ancestry proportions, while they are negatively correlated with the proportion of African, European and East Asian ancestries (**Figure S9C**). This observation is seen across length classes of ROHs, though more attenuated for long ROHs (**Figure 3C**). We reasoned that this ancestry effect across length classes is likely reflecting the long-term endogamous marriages and recent consanguinity associated with the ME-like ancestry relative to admixture component of other ancestries.

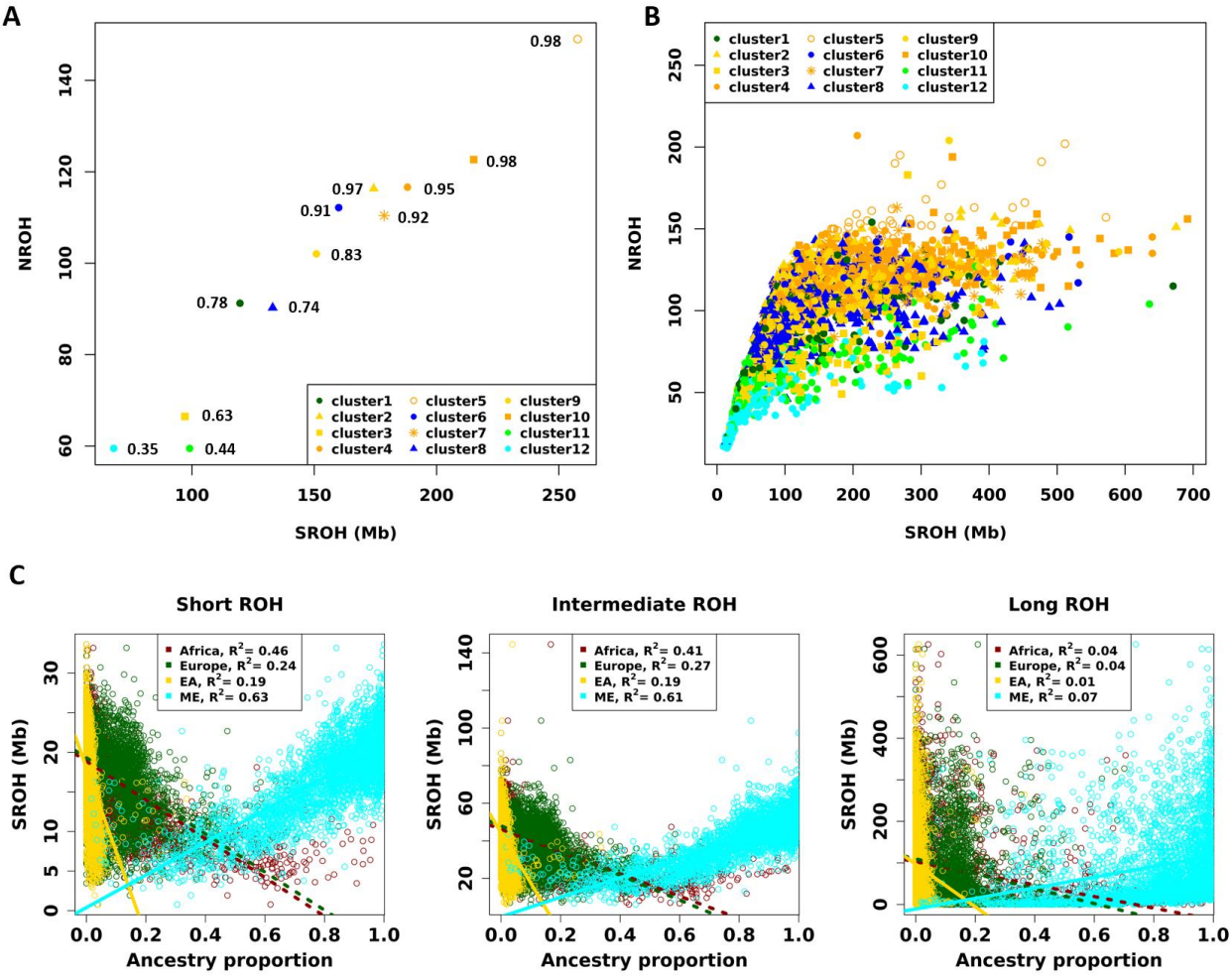


Figure 3. Runs of homozygosity in Saudi Arabians. (A) Average total length and number of ROH per cluster. The numbers next to the symbol represents the mean ME-like ancestry proportion. (B) Total length and number of ROH per individual across the Saudi Arabian cohort. For (A) and (B), symbols are colored by the geographical region associated with each cluster (**Figure 1D**). (C) Total length of ROH vs ancestry proportion per individual stratified by three length classes of ROHs. ROH – Runs of homozygosity, ME - Middle Eastern, EA – East Asia.

We leveraged the dense marker information from the 302 WGS individuals to reconstruct genome-wide genealogies and infer the population size trajectories within the Saudi social substructure (**Figure 4**). Consistent with the out-of-Africa event, all clusters experienced and subsequently recovered from a decline in the effective population size (Ne) about 100 kya. All

Saudi clusters reached a local maximum Ne around 9 -10 kya, a period consistent with the early Holocene Wet Phase / Holocene Humid Period, characterized by wet conditions which resulted in expansion of lakes and rivers and extensive grasslands [60]. Following the Holocene period, populations in the Arabian Peninsula experienced another bottleneck dating around 6 - 7 kya, along with divergence among sub-clusters. This period coincided with the Arabian aridification, which is responsible for the desert conditions in most of the Arabia as we know it today [6,60]. Clusters with less ME-like ancestries and stronger signature of admixture (such as clusters 12, 11, 8, & 3), showed less severe decline in Ne compared to those that have high ME-like ancestry. Cluster5 in particular, showed the most severe bottleneck and remained low in Ne in the recent times, consistent with long-term isolation. Cluster5 appears to resemble the pattern of the tribe labelled as T25 in a previous study [20]: both originated from the Western region showing the highest level of inbreeding within the respective study. T25 is said to have been subjected to strict intratribal marriages. Such social practices can indeed result in persistent small Ne as observed here, as well as our observed pattern in ROH (**Figure 3**).

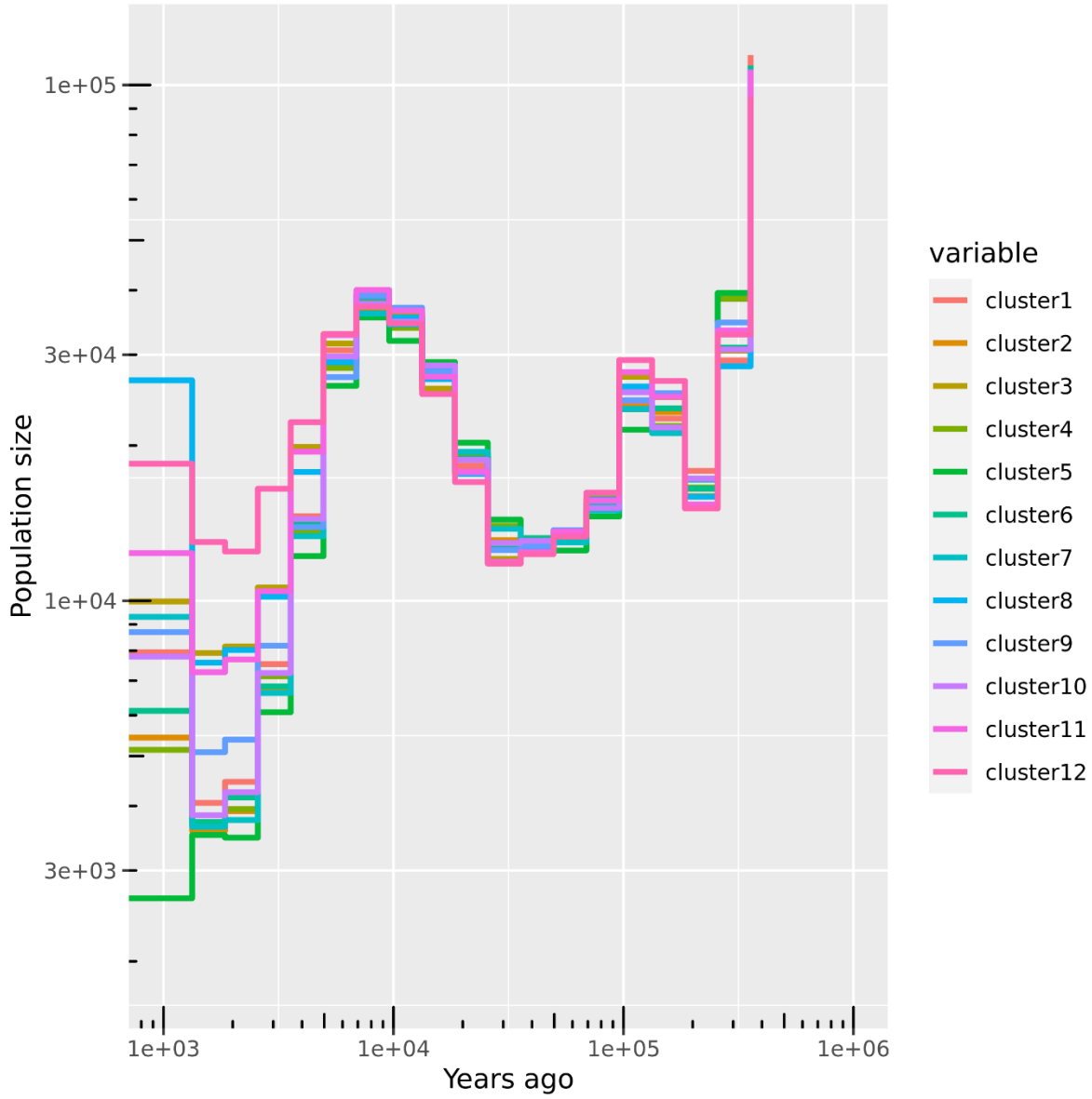


Figure 4. Population size trajectories between the Saudi Arabian sub-clusters. Effective population sizes were computed from genealogical trees using RELATE (see **Methods**). The number of samples per cluster used for the estimates can be found in **Table S1**.

Saudi's social structure does not impact imputation accuracy but lack of reference representation does

Using the Trans-Omics for Precision Medicine (TOPMed) reference panel [61,62], we imputed the genotypes based on Saudi array data and compared the imputation accuracy of the Saudi to the Europeans from the United Kingdom 10,000 Genomes (UK10K) project matched by sample size and SNP content to further evaluate the impact population history may have on haplotypic pattern of variation and implications for genetic epidemiology studies in Saudi today. Unsurprisingly, the imputation accuracy was lower for Saudi compared to Europeans across all MAF bins (**Figure S10A**). This is consistent with what was reported in Cahoon et al., [63] which showed Saudi Arabia among the populations with the lowest imputation accuracy when compared to Europeans and populations within North America. Across Saudi sub-clusters, the imputation accuracies were quite similar, except for a slight difference in cluster12 which had lower imputation accuracy of common variants (**Figure S10B**). This observation is consistent with Cluster12 showing elevated shared drift with Africans from the Southern and/or Eastern region (**Figures 2 and S7**), which may not be well-represented in the TOPMed.

Allelic architecture of Saudi Arabians

Genetic variation in Saudi Arabian WGS data

Having investigated extensively the population structure, admixture and demographic history and their impact on ROHs in the Saudi, we then leveraged the WGS data from 302 Saudi individuals to investigate the consequence of population history on the pattern of genetic variation in Saudi. A total of 25,488,981 autosomal variants were called and retained after quality control (QC) (**Methods**), of which 2,459,950 (9.7%) variants were not previously identified in gnomAD v4.1 [64,65] and thus are potentially novel or Saudi-specific. We refer to these variants as the “previously unknown variants”. As expected, the previously unknown variants are highly enriched with rare alleles (83% of them are singletons in our dataset, compared to 32% singletons among the known variants; **Figure S11**). Of all variants, 63% have MAF < 1%. Although there are some Middle Eastern individuals in gnomAD v4, they are

proportionally underrepresented in this global dataset (2,884 exomes among 730,947 total, 147 genomes among 76,215 total), resulting in a large number of Saudi Arabian variants missing from the database (although pipeline differences may explain some of the missing variants). Previously, Almarri et al., [6] found that of 23.1 million single nucleotide variants identified in 147 Arabian and Levantine individuals, 4.8 million (20.8%) were not found in the Human Genome Diversity Project (HGDP-CEPH) global dataset. Taken together, both our and Almarri et al., studies showed that variation from the Arabian Peninsula are not yet well captured. Thus, genetic association studies will be limited to only the common variation, which we have also shown to be currently sub-optimally imputed (**Figure S10**). Here we provide all the 25,488,981 variants that remained after QC and their allele frequencies, see the "**Data availability**" section for access**.**

We compared allelic frequency spectra and allelic homozygosity in the Saudi Arabians (all WGS individuals) to the Middle Eastern population in gnomAD (gnomAD-MID). The two have relatively similar patterns in the genome-wide alternative allele frequency spectra though Saudi had proportionally slightly fewer common variants (**Figure S12A**). The allele frequencies are highly concordant (r = 0.98) between the two populations (**Figure S12B**), but Saudi Arabians have approximately 2x more homozygous genotypes than gnomAD-MID (e.g. an average of 20% vs. 10% of the genotypes are homozygous for variants with alternative allele frequency > 5% in Saudi and gnomAD-MID, respectively; **Figure S12A** and **S12C**). The higher proportion of inbreeding suggests that the Saudi and the gnomAD-MID population are not reflective of the same underlying populations. However, because the frequency spectra and correlation of allele frequencies are highly similar (**Figure S12**), we thus utilize both samples to compare the pattern of variation with gnomAD African/African Americans (gnomAD-AFR) and non-Finnish Europeans (gnomAD-EUR) below to better understand the impact of the unique history in the Arabian Peninsula on its current pattern of variation.

Distribution of functionally deleterious variants

We annotated the variants using three different annotation tools: VEP (v.110) (McLaren et al., 2016), AlphaMissense (Cheng et al., 2023), and Genomic Pre-trained Network (GPN) (Benegas et al., 2023). AlphaMissense predicts the pathogenicity of missense variants [66] while GPN predicts the deleteriousness for both coding and non-coding variants. The distribution of the variants by functional classes are shown in **Table S18**. We first examined the set of previously unknown variants, relying on AlphaMissense and VEP only as GPN is precomputed only for variants found in gnomAD [67]. In addition to being enriched with rarer alleles, proportionally more unknown variants (19.63%) were annotated to be deleterious than known ones found in gnomAD (7.2%). This Implies that the previously unknown variants are not just sequencing errors distributed randomly across the genome, but are enriched for rare variants of functional relevance in the Saudi that are maintained or have not been purged from the population.

Non-Africans are expected to have more deleterious alleles due to the relaxation of purifying selection during the Out-of-Africa (OOA) bottleneck as well as the introduction of new deleterious mutations during population expansion [30,68,69]. Despite some opposing reports on this hypothesis [70,71], there has been empirical evidence in isolated populations having an excess of functionally deleterious alleles [72–74]. In addition to the OOA bottleneck, Saudi has a deep culture of endogamy and consanguinity, and these demographic factors are known to potentially increase the burden of deleterious alleles in a population due to decreased efficacy of purifying selection [33]. Here we investigated the allelic architecture of functionally deleterious alleles in the Saudi population, compared to other continental populations from gnomAD. Compared to gnomAD-AFR individuals, the Saudi tend to show proportionally more deleterious alleles than those annotated to be benign or neutral across algorithms (**Figure 5A**), particularly for variants up to ~5% frequency. Overall, relative to gnomAD-AFR, between the 0.5 - 5% frequency, we found a 13% proportional increase of deleterious (likely pathogenic) alleles annotated by AlphaMissense in the Saudi Arabians compared to 7% proportional decrease of the benign alleles ($P < 0.01$; **Figure 5A**). When annotated by VEP and GPN, at the same frequency range, we observed a consistent pattern i.e. a 3% proportional increase in loss of function variants in the Saudi Arabians compared to 10% proportional decrease in neutral (synonymous) ones ($P < 0.01$) by VEP, and an 11% proportional increase in the first percentile of

alleles by deleteriousness compared to 3% proportional decrease in the 99th percentile (e.g. the most likely neutral) of alleles when annotated by GPN (**Figure 5A**).

This pattern of enrichment for functionally deleterious alleles is also qualitatively observed when comparing the exome samples from gnomAD-MID to gnomAD-AFR (**Figure S13**), taking advantage of the larger sample size for gnomAD-MID exomes and the same data processing pipeline in gnomAD. The pattern is also qualitatively observed when comparing Saudi to gnomAD-EUR, though the difference may be more attenuated in some allele frequency bins (e.g. for AlphaMissense annotation; **Figure S14**). The less significant finding when comparing Saudi to gnomAD-EUR is probably because Europeans also showed proportionally more deleterious than neutral alleles across all frequency bins, as previously reported [68,75] and replicated here (**Figure S15**).

We also compared the enrichment of deleterious alleles between Saudi sub-clusters. Because of the smaller number of individuals within each cluster having WGS data (**Table S1**), we grouped the clusters into two groups: groupA which contained clusters with greater inbreeding and lower effective population sizes (clusters 2, 4, 5, 6, 9, and 10), and groupB which has less inbreeding and higher effective population sizes (clusters 12, 11, 3, and 8). We left out cluster1 from this analysis as it tends to fall in the middle of the two groups. GroupB had generally greater number of variants compared to groupA (**Figure S16**), consistent with its higher genetic diversity and less inbreeding. GroupA, with greater inbreeding and lower effective population sizes, showed greater enrichment of deleterious alleles (**Figure 5B**).

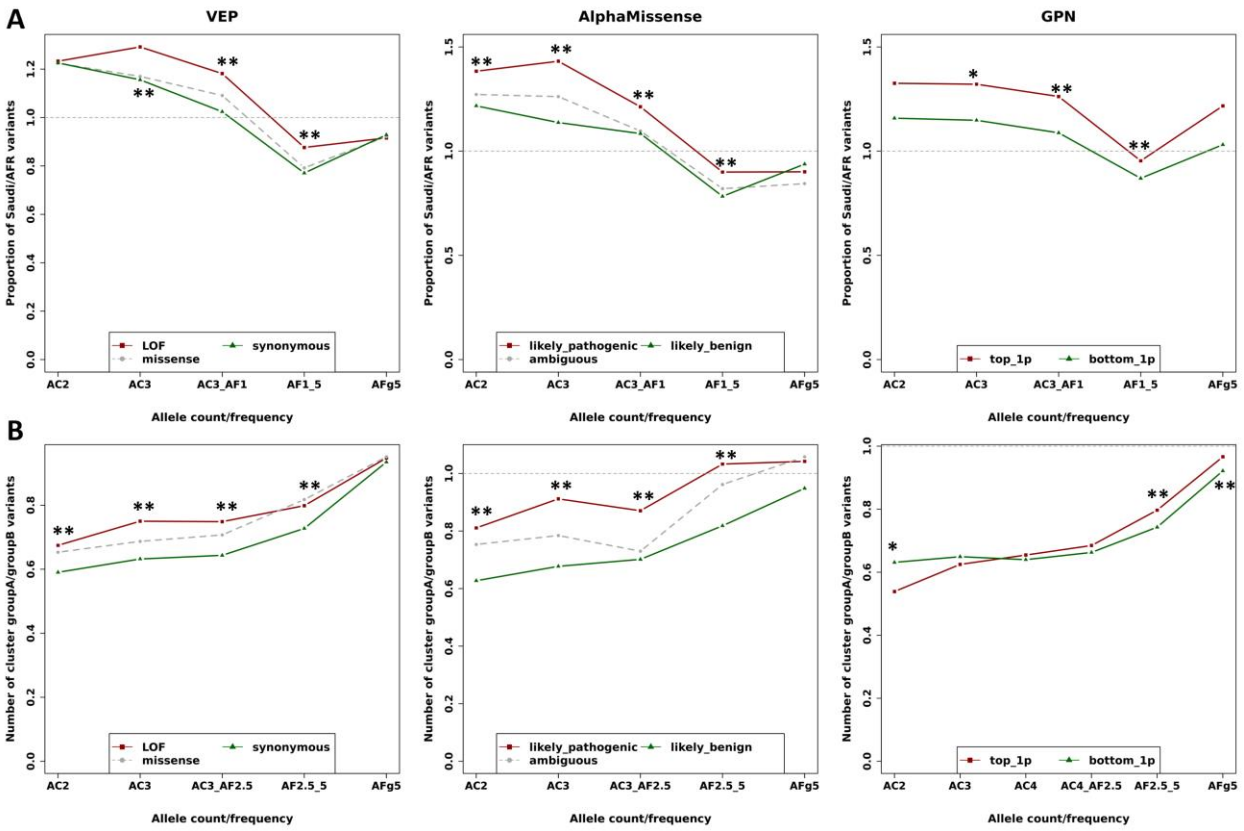


Figure 5. Distribution of minor allele frequency across functional classes. (A) Ratio of Saudi to gnomAD-AFR variants. The sample size of gnomAD-AFR is based on downsampling to Saudi sample size, n = 302. (B) Ratio of Saudi cluster groupA to cluster groupB variants. The sample size of cluster groupB is based on downsampling to groupA sample size, n = 124. Variant functional consequences were annotated based on VEP (loss-of-function, missense, or synonymous variants), AlphaMissense (likely pathogenic, likely benign, and ambiguous), and GPN. AC and AF refer to allele count and allele frequency, respectively. AFg5 refers to allele frequency greater than 5%. Top_1p refers to variants with the top 1% of GPN scores (more deleterious) and Bottom_1p refers to variants with the bottom 1% of GPN scores (more neutral). AFR denotes the gnomAD-AFR sample. LOF refers to Loss of function. ** and * denote frequency bins with significant difference between the most deleterious (red) and most neutral (green) through bootstrapping at $p < 0.01$ and < 0.05, respectively.

DISCUSSION

Scholars have noted the complexity of diverse histories in shaping the genetic architecture of Arabian Peninsula populations, and called for better characterization of each population for better understanding of their genetics and health [35,76]. On one hand, being situated

509 geographically at the crossroads between Africa and Eurasia, which facilitates intercontinental
510 interactions, is expected to increase heterozygosity and genetic diversity in the AP populations.
511 On the other hand, Saudi Arabian culture is rooted in endogamous practices which increases
512 homozygosity which can result in health consequences [22,25]. Here we elucidated the fine-scale
513 population structure of Saudi Arabian population using 3,252 genotyped and 302 WGS
514 individuals from various geographic regions within the country and investigated the genetic
515 variation within the social structure and impact of the demographic histories on the population
516 health.

517 Our analyses concur with previous findings about the presence of sub-population structure
518 within Saudi Arabia [20,77], with twelve distinct genetic sub-clusters being identified in our data.
519 We chose to infer the sub-clusters based on genetic similarities and use the resulting sub-
520 clusters as units of analysis throughout the study. We note that clustering by genetic similarity
521 in an admixed population could misrepresent the population structure in the dataset, such as
522 when multiple clusters of Arabian origin are combined into a single cluster due to sharing a
523 common admixing source (e.g. African). However, we elected for this approach in part because
524 of limited tribal affiliation information at the individual level due to privacy concerns, thus we
525 could not rely on self-reported tribes as units of analysis. Moreover, ancestry-specific
526 approaches to infer population structure [78,79] are limited to situations where the admixing
527 ancestries are divergent and that references for the ancestral populations are available, both of
528 which are understudied or unavailable for Saudi Arabians. While our observation of population
529 structure should be re-examined in the future with more self-reported demographic
530 information or with improved methodologies to exclude the impact of admixture, we also note
531 that our clustering is not driven solely by non-Arabian admixture. For instance, both cluster12
532 and cluster3 exhibit strong levels of African admixture (**Figure 1C; Figure 2A - B**), but they also
533 showed affinity to different Bahraini aDNA references with varying levels of Levantine-related
534 ancestry (**Figure 2A**). Therefore, in this case, level of African admixture is not the only reason
535 that separated clusters 12 and 3 from the rest of the Saudi Arabian with less admixture.
536 Furthermore, the fine-scale structure in the Saudi Arabian population, brought on at least in
537 part by the social practices, has previously been reported when studying self-identified

538 indigenous tribes [20] and we do observe some similarities of such structure with our clusters.
539 Proceeding with the genetic sub-clusters defined, we then continue to elucidate several key
540 features in the pattern of genetic variation in Saudi Arabia.

541 While geographic proximities increase the chances for gene flow between individuals around
542 same geographic areas, the strong consanguineous and endogamous culture in Saudi Arabia is
543 expected to limit such interactions, resulting in distinct sub-clusters residing in relatively close
544 proximities that we and others have observed [20]. We also note that, while recent and/or
545 ongoing admixture may be taking place between the Saudi Arabians and modern-day Africans,
546 Europeans and Central & South Asia, the spread of such genetic pool is likely restricted by the
547 endogamous and consanguineous practices as previously suggested for the Emiratis [35]. Both
548 endogamy and consanguinity increase the burden of ROH (SROH) which increases health risks.
549 In El-Mouzan et al., [80], the highest rate of consanguineous marriages was found in Madinah
550 from the Western region of the country. Similarly, our results show footprints of inbreeding in
551 all the sub-clusters from the Western region (clusters 4, 5, 7, and 10) as demonstrated by the
552 longest ROHs, lowered Ne, and less evidence of admixture. On the other hand, the sub-groups
553 that intermarry benefit from the increased effective population sizes and genetic diversity,
554 especially those with elevated African admixture. The sub-clusters with elevated admixture in
555 our study also show reduced ROH numbers and sizes.

556 Consistent with the Arab slave trade and Islamic expansion in the 7th (1,300 – 1,400 years ago)
557 century, almost all of the Saudi clusters show a recovery in effective population sizes from the
558 Arabian aridification (about 6 - 7 kya) bottleneck, at varying ranges, except for the isolated sub-
559 cluster from the Western region (cluster5). Whilst the admixture dates for Saudi cluster5 qpAdm
560 model of Ethiopia_4500BP + Egypt (1120±330 years ago) and Ethiopia_4500BP + MH2_MH1_LT
561 (1242 ± 528 years ago) align with the 7th century Islamic expansion we note that historical
562 documentation of significant North African to Arabian migrations during this period remains
563 limited and thus, the possible conduit for Northern/Northeastern African ancestry to Saudi
564 cluster5 remains inconclusive. The increase in the effective population sizes from the recent
565 times is more for sub-clusters with higher African admixtures. Previous studies have reported

dominating African admixture source in Arabia originating from Bantu speakers from the East (Kenya) or South Africa, dating from 400 to 1754 years ago [6,13,81], which is consistent with the Arab trade slave expansion in the 7th (1,300 – 1,400 years ago) century. In a recent study [20], the Saudi tribes with highest African admixture were results of recent admixture events, as recently as 11 generations ago. This may be a sign of continuous admixture beyond the Arab trade slave expansion. Indeed, we observe a recent (358 ± 27 years ago) estimated timing of Saudi cluster12 ancestry formation of Ethiopia 4500BP + MH1-MH2 LT. This recent timing, coupled with distinct Central, Eastern, and Southern African genetic affinity, suggests a possible connection to 17th - 20th century Red Sea and trans-Saharan slave trades, as opposed to the older North African Islamic expansion, during which East African regions, including Ethiopia and Eritrea (historically referred to as Abyssinia), Somalia, and Sudan, were significant sources of enslaved individuals [82]. Additionally, Central African regions extending into present-day Chad and the Congo Basin contributed to this population due to extensive trade networks.

The impact of the OOA bottleneck and population size histories on the allelic architecture of deleterious alleles in non-African populations has been a matter of debate [30–32,68–71,73]. We observed an abundance of rare and low frequency (AF = 0.5 – 5%) deleterious alleles in Saudi Arabians when compared to gnomAD-AFR. This enrichment in deleterious alleles can be explained to some extent by the demographic history of the Saudi Arabians beyond the OOA bottleneck event. First, the high percentage of previously unknown deleterious mutations that we found in sequencing could be driven by consanguinity/inbreeding. Second, although consanguinity and endogamy could in principle expose deleterious alleles to negative selection through genetic purging, purging is less effective when effective population sizes are small thereby the deleterious alleles may drift to high frequencies and even become fixed [33,75], particularly if population isolation followed a bottleneck [31,72,73,83]. Saudi Arabia is not a completely isolated population, but there is reproductive isolation due to their social practices. Taken separately, the subgroups with high prevalence of endogamy are indeed enriched for deleterious alleles than those that have high levels of admixture (**Figure 5B**). Our results are consistent with other empirical studies that show enrichment of deleterious in populations with

different demographic histories following population bottleneck [72,75] and further confirms no evidence of genetic purging in the Saudi Arabians as previously reported [26].

Overall, our results shows that Saudi's population history impacts its pattern of genetic variation with potential consequences to the population health. The legacy of endogamy and consanguinity in the population poses health risks and the frequency of this practice has not shown signs of decline [84,85]. There have been initiatives to raise public awareness on the health risks of close relative marriages in Saudi Arabia and other Middle East countries which include mandatory premarital health screening for recessive illnesses and genetic counselling [22,23,86,87]. Even though the mandatory premarital screening in Saudi Arabia has so far not been very successful in discouraging or preventing at-risk marriages, it has fostered a more informed pre-birth decisions, reducing the prevalence of children born with the health complications through altering some of the cultural behaviors including adoption of prenatal detection and therapeutic abortion [14,85,87]. With increasing public education and awareness on risk factors associated with endogamous and consanguineous unions, its prevalence may change in the future, especially because the consanguineous marriages are generally most prevalent in poor, rural and least educated societies of Arabia compared to urbanized and more educated counterparts [88].

We note a common issue with regards to the under-representation of Arabian countries in global cohorts which is also raised by others [6,26,89]. For example, for the genomes in gnomAD database, only 0.2% of samples are of Middle Eastern origins, compared to 44.6% and 27.3% as Europeans and Africans, respectively. Even with the newly increased exome data, the Middle Easterners only make up 0.38% of gnomAD, compared to 4.65% Africans and 77.07% Europeans. As a result, the human genetics field in general is missing variants that are enriched to Saudi Arabia and Middle East. It has also been suggested that GME populations tend to harbor more variants unique to the region [26] e.g. 28% in the Qatar [90]. The under-representation accentuates the understudying of regionally enriched alleles, and contributes to reduced accuracy of polygenic prediction models when applied to Arabic populations [91] and lower imputation accuracy when using state-of-the-art reference panel like TOPMed [63] (**Figure S10**).

This further highlights the need to sequence diverse and unique populations and include them in large global genetic data platforms such as gnomAD, TOPMed and others.

METHODS

Data collection, processing and quality control

For all studied samples, written informed consent was obtained from each participant. The studies were approved under the Saudi Genome Project by the Institutional Review Board at King Abdulaziz City for Science and Technology and King Fahad Medical City. In compliance with Saudi privacy legislation and the protection of human subject confidentiality, the sharing of raw genotyping and clinical data is restricted. Access to this data requires prior approval from the Saudi National Bioethics Committee.

Array data

Sample collection, genotyping and quality control. A total of 3,752 samples were collected in Saudi Arabia between the years 2017 - 2020 as control individuals for various projects, such as the GenOMICC International project and covid19 host genetics consortium [92] studies. Individuals were genotyped on the Axiom Genome-wide CEU 1 Array including customized variants following the manufacturer's specifications for sample preparation, including whole genome amplification, fragmentation, denaturation, and hybridization. Genome-wide SNP genotyping was performed using the automated, high-throughput GeneTitan system from Affymetrix.

We filtered individuals with sample call rates < 0.9 using PLINK v1.9 [93,94] on each plate individually before merging the autosomal SNPs across the different plates, resulting in a merged set of 757,790 SNPs. We removed duplicates and non-biallelic variants, retaining 703,986 SNPs. We then filtered SNPs with greater than 10% missing rate and SNPs that did not

pass Hardy Weinberg Equilibrium (HWE) test ($P < 10^{-6}$) using PLINK, resulting in a total of 606,349 SNPs for analysis. We lifted over the genomic coordinates from human reference genome hg19 to hg38. We phased the data using Beagle v5.2 [95].

Removing close relatives and filtering out outliers. Using the 3,752 Saudi samples and 606,349 SNPs, we pruned the dataset by linkage disequilibrium (LD) (using the command --*indep—pairwise 50 5 0.8* in PLINK), resulting in 547,307 SNPs to estimate individuals' relatedness using King v2.2.5 [96]. We removed twins (or duplicated individuals) as well as first degree relatives, retaining 3,403 samples. Furthermore, we performed PCA and performed two iterations of outlier (defined as being > 6 standard deviation (SD) away from the mean in any of the first 10 PCs), resulting in 3,352 samples left for further analyses.

Defining samples' tribal affiliation and imputing missing tribal information. We aimed to use available demographic information, i.e. tribal affiliations, in validating and interpreting the results of clustering based on genetic data. However, 82% of the individuals in our data (2,740 of the 3,352) did not have self-reported tribal information. We thus imputed such information using the software HARE (harmonized ancestry and race/ethnicity) package [97] based on the available Self-identified Race/Ethnicity (SIRE) tribal information of 612 individuals. SIRE in our data were derived from either self-report or individual's family name that is presumed to reflect their tribal affiliation (**Table S1**). The HARE package combines genetically inferred structure based on PCA with available SIRE information to train a support vector machine (SVM) classifier that could correct for potentially mislabeled SIRE and predict the race/ethnicity, in this case tribal label, for those individuals missing SIRE. We used the HARE to impute tribal information of the samples missing a SIRE label in our dataset using the first 30 PCs as the input data. We used the highest predicted membership probability (L_1, see Fang et., [97] for more details) labels to aid in the interpretation of the population sub-clusters that we infer from genetic data.

Whole genome sequencing (WGS) data

Sequencing information and processing. In addition to the genotyped samples, 349 samples were whole-genome sequenced (WGS) to a targeted depth of 30x. The samples were prepared following the Illumina's TruSeq Nano sample preparation protocol and sequenced on an Illumina HiSeq X-ten machine. The raw sequences were aligned against the human reference genome GRCh38 using the Burrows-Wheeler Aligner (BWA) version 0.7.10 [98]. Picard tools version 1.117 was used to mark duplicates [99]. All sample preparation, sequencing, sequence alignment, pre-processing, quality control before calling of variants and BAM file augmentation were performed by deCODE genetics (https://www.decode.com), and a more detailed information on these steps is documented in Jónsson et al., [100].

Variant calling and filtering. We merged the gVCFs of the 349 samples using CombineGVCFs in GATK [99] and subsequently performed a joint genotyping calling using GenotypeGVCFs. We performed variant quality score calibration (VQRS) on the combined samples using VariantRecalibrator and ApplyVQRS in GATK [101]. We supplied the homo sapiens reference assembly 38 (Homo_sapiens_assembly38.fasta) and used the following resources: HapMap III variants were used as training and truth sets with prior priority of 15, 1000G omni2.5 sites were used as training set with prior priority of 12, 1000G phase1 high confidence SNPs was used as training set with prior priority of 10 and the dbSNP138 as known SNPs with prior probability of 2. For the annotations, we included the QD, MQ, MQRankSum, ReadPosRankSum FS and SOR. We used 99% sensitivity level to filter the SNPs.

Quality control on samples and SNPs. All 349 samples had missing genotyping rate < 10%. We excluded 302,640 SNPs with missing rate > 10% and 53,981 SNPs based on HWE threshold ($P < 10^{6}$), leaving 26,781,476 SNPs. We removed non-biallelic sites which left 26,408,559 variants. Further filtering was applied on specific downstream analyses when appropriate. To exclude outliers in our samples, we merged the 349 WGS samples with our array data and the HGDP dataset at segregating SNPs shared across all datasets. A principal component analysis (PCA) was performed using PLINK and we used HARE to impute missing self-reported individual nationalities (e.g. self-identified nationality as Saudi or not). We excluded 8 samples which were

not imputed as a Saudi. We then removed monomorphic sites which were introduced by calling the variants including these potentially non-Saudi samples, leaving 25,488,981 variants.

We filtered samples based on relatedness using King software v2.2.5 [96]. For estimating the relatedness, we randomly sampled 550,000 SNPs with minor allele frequency > 1% after LD pruning (*--indep-pairwise 50 5 0.5* using PLINK**)** to estimate the relatedness. We removed 37 twins/duplicates and first-degree relatives. Using the PCA, we further removed 2 samples that appeared as extreme outliers (> 6 SD on any of the first 10 PCs), leaving 302 samples. Haplotype phasing was performed on the remaining 302 samples and 25,488,981 variants using Eagle v2.4.1 [102].

Annotation of variants

We annotated the variants using the popular VEP (v.110) [103] as well as two recently published annotation tools, AlphaMissense [104] and Genomic Pre-trained Network (GPN) [67]. The AlphaMissense only annotates missense variants and has three functional classes, “likely pathogenic”, “ambiguous” and “likely benign”. The GPN annotates all genomic variants and assign a deleteriousness score to each variant in gnomAD (v3). We downloaded the pre-computed scores from https://huggingface.co/datasets/songlab/gnomad/resolve/main/test.parquet, accessed 2/9/2024.

Merging of Saudi whole-genome-sequence data with ancient genomes

We downloaded the Allen Ancient DNA Resource (AADR) v.54.1 Eigenstrat files which are genotyped according to hg19 coordinates. We were kindly provided the Bahrain aDNA [12] genome bed files by Rui Leite Portela Martiniano, which were originally mapped to GRCh38. To have all our genotype files on a consistent reference genome, we mapped the Saudi whole-

genome-sequenced data and ancient Bahrain samples back to human reference genome hg37 using liftOver. We then filtered the variants though PLINK 1.9-beta7 using parameters *--geno 0 --snps-only --make-bed --allow-no-sex*. Prior to merging the Saudi WGS, AADR, and Bahrain datasets we filtered mistyped SNPs where the rsIDs are shared between the datasets, but the AADR reference allele does not match either the reference or alternate allele (n=520). The Saudi WGS and Bahrain aDNA datasets were then converted from plink to packedancestrymap format through Eigensoft convert function with parameter *familynames: NO*. Finally, we merged the Saudi WGS and AADR datasets with the Eigensoft mergeit function with parameters *strandcheck: YES*. The mergeit program merges two data sets into a third, which has the union of the individuals and the intersection of the SNPs in the first two. We first merged the AADR and Saudi WGS datasets. The merging of the AADR and Saudi datasets resulted in the filtering of 14,239 SNPs due to A/T or C/G strand checks and 91 SNPs due to allele mismatch. In addition, there were 770,115 genotype strand flips with the final dataset consisting of 1,032,250 retained SNPs. We then merged the Bahrain aDNA resulting in a final packedancestrymap genotype file of 1,030,352 SNPs.

Data analyses

We used the larger collection of Saudi genotyped samples to investigate the genetic substructure and historic admixtures of the population. We then utilized the high-density genome-wide marker information from the WGS data to investigate differences in genetic ancestries with aDNA, population size trajectories, and allelic architecture of functional variants within the social structure of the Saudi population.

Evaluation of population structure

We merged the fully filtered array and WGS datasets, based on segregating markers. We performed PCA followed by UMAP [105] to combine the first 10 PCs and reduced them into two-dimensions in order to explore the population structure. Based on the UMAP results, we

assigned individuals to subpopulations using K-means clustering from the R package *stats* [106]. To determine the optimal number of K clusters, we used the Average Silhouette Width (ASW) (**Figure S17;** [107]) which is a popular and trusted method to produce quality clustering [108]. The ASW uses values between -1 and 1 to measure how similar/dissimilar is an object to others within its cluster as well as objects in different clusters, with higher numbers representing a better fit and appropriateness of clustering. Likewise, a high ASW value corresponds to an optimal number of K clusters for partitioning a particular set of objects [107]. We validated these clustering by evaluating the concurrence between the clusters and the tribal region assignments. We used these clusters as representative of the social structure and also used them in the whole genome sequencing samples to evaluate patterns of genetic diversity within the Saudi population.

Analysis of ancestry components. We conducted the unsupervised admixture analysis using ADMIXTURE software v1.3 [109]. We conducted 10 independent runs of admixture analysis for each K and retained the run with maximum likelihood. We used the cross-validation procedure, implemented in the program, to identify the best number of ancestral populations K which fits our data.

Evaluating patterns of admixture. To further test for the presence of admixture within the identified clusters, we performed supervised admixture analysis using the f_3-statistics from the ADMIXTOOLS 2 package v2.0.4 [110]. We computed the f_3-statistics using the Saudi clusters as targets and using all pairs of populations in the HDGP data [34] as source populations i.e. f_3(Saudi cluster; HGDP population 1, HGDP population 2).

We also used the outgroup f_3-statistics to investigate the degree of shared drift between Saudi clusters and the HGDP populations. For this statistic, the African San, Mbuti or Yoruba are often considered as outgroups for investigation of non-African populations. However, HGDP African populations have showed to be highly admixed with some of the Saudi clusters, and the outgroup should be close enough but should not be part of the ingroups [111]. We first used f_4-statistics of the form f_4(HGDP population, HGDP population; Saudi cluster, Saudi cluster) to

776 determine a suitable outgroup for the Saudi clusters from the HGDP reference populations, as
777 recommended by Pattersons et al., [111]. We found that nearly all HGDP population combinations
778 showed positive gene flow between the HGDP population and either one or both of the Saudi
779 clusters, thereby violating the outgroup assumption. The HGDP populations that did not violate
780 the outgroup assumption in this test were eight East Asian populations: Han, Miao, Japanese,
781 Tujia, Yi, Hezhen, She and Naxi. We then used Han as an outgroup for the outgroup f_3-statistics
782 in the form: f_3(outgroup; population1, Saudi cluster), whereby population1 was an HGDP
783 population or another Saudi cluster except the target.

784 **Saudi and ancient genome analyses.** To assess the genetic affinity of Saudi clusters to present-
785 day and ancient African groups, we computed the f_3-statistic of form f_3(Han; African groups,
786 Saudi clusters) using the ADMIXTOOLS 2 package v2.0.4 [110,112]. We removed individuals that
787 were indicated on the AADR metadata as relatives, contaminated, duplicated, or have low
788 coverage. We tested the below allele sharing pattern using f_4-statistics in ADMIXTOOLS 2 with
789 parameters *f4mode = TRUE, afprod = TRUE, allsnps = TRUE*. To estimate 'Basal Eurasian'
790 ancestry in the Saudi clusters using a selection of eight ancient African groups:
791 MAR_Taforalt_EpiP [49], Egypt_ThirdIntermediatePeriod [51], Ethiopia_4500BP [54,113],
792 Kenya_Nyarindi_LSA_Kansyore [114], Tanzania_Zanzibar_1300BP [115], Malawi_Fingira_LSA_6000BP
793 [54], South_Africa_400BP.SG [55], and Cameroon_SMA [54] in the f_4-statistic of form f_4(Saudi cluster,
794 Han.DG [34,116,117] Ust-Ishim, African group). The North African Epipalaeolithic Moroccan
795 Iberomaurusian Taforalt group (Taforalt EpiP) represents the best proxy of Basal Eurasian
796 ancestry [7], exhibiting genetic connections to both early Holocene Near Easterners, such as
797 Levantine Epipaleolithic Natufians (Natufian EpiP), and sub-Saharan Africans. Thus, to further
798 assess the genetic affinity of Saudi clusters to African ancestry relative to the shared ancestry of
799 North African Upper Paleolithic Taforalt Moroccan and Epipaleolithic Natufian Levantine groups,
800 we ran the f_4-statistic of form f_4(Saudi cluster, Upper Paleolithic Taforalt; African group,
801 Epipaleolithic Natufian). Finally, to assess the relative shared drift between Epipaleolithic
802 Levantine Natufian [8] and Neolithic Central Zagros [8,118] and CHG [119] ancestries, we used the f_4-
803 statistic of the form f_4(Saudi cluster, Yoruba.DG [34,120]; Ganj Dareh N/CHG, Natufian EpiP).

We employed replacement qpAdm [112] with parameters *allsnps=TRUE and fudge_twice=TRUE* to model the ancestry for each of the Saudi clusters as it is partitioned in ancient groups across four broad and approximate periods. For each period we kept the following core fixed right group set of populations: Mbuti.DG [121], Papuan.DG [120], Russia_Ust_Ishim.DG [122,123], Russia_MA1_HG.SG [124], Russia_Kostenki14 [125], WHG [10,119,125,126], CHG [119], EHG [127], Turkey_Epipaleolithic [128], Iran_GanjDareh_N [8], and ISR_Natufian_EpiP [8]. For each of the four-time bins we modeled one to five source models, cycling through the source populations to form each qpAdm model. In evaluating the qpAdm models, we preferentially selected the model with the least number of sources with the largest p-value (with a plausibility threshold cut-off of 0.01 and admixture weights between 0 and 1), iteratively evaluating more complex models.

Pre-Pottery Neolithic to Neolithic Sources: Italy_Sardinia_N [38,40], Levant_PPN [8,129], Mesopotamia_PPN [130], Anatolia_Marmara_Barcin_N [127,129], Turkey_Catalhoyuk_N_Ceramic.SG [131], MAR_Taforalt_EpiP [49], Armenia_Aknashen_N [129], Armenia_MasisBlur_N . **Chalcolithic to Bronze Age Sources**: Iran_C_SehGabi [8], Turkey_TellKurdu_EC, Turkey_C [8,129], Israel_C [132] , Armenia_C [8], Steppe_Eneolithic [133], and Ethiopia_4500BP [54,113]. **Bronze Age Sources:** Ethiopia_4500BP, Italy_Sardinia_EBA [38,40], Mesopotamia_LBA [129], Israel_MLBA [134], Jordan_LBA [134], Lebanon_MBA.SG [135], Turkey_EBA [129], Armenia_EBA_KuraAraxes [129,133], Armenia_MBA [8,129], and Germany_BellBeaker [127,136]. **Bronze Age to Historical Sources**: Ethiopia_4500BP, Italy_Sardinia_LBA [40], Germany_BellBeaker [127,136], Iran_Hasanlu_IA [129], Turkey_IA [129], Egypt_ThirdIntermediatePeriod, AS_EMT [12], MH2_MH1_LT [12], MH3_LT [12], Hungary_IA_LaTene [137], Israel_Ashkelon_IA2 [138], and Jordan_LBA_IA [134].

We sought to date the formation of the plausible qpAdm models with DATES v4010 [46] using parameters *binsize: 0.001, maxdis: 1.0, qbin: 10, runfit: YES, qbin: 10, runfit: YES, afffit: YES, lovalfit: 0.45, samecoeffs: NO, and jackknife: YES*. For each of the plausible qpAdm models, we ran the combination of sources through DATES, evaluating models with a normalized root mean standard deviation of < 0.7 & Z-score > 2 as well fitting. To obtain admixture dates in calendar years we used a generation time of 28 years [139].

Runs of homozygosity. ROH are continuous segments of homozygous genotypes inherited from common ancestor [56]. Following Choudhury et al., [140] we used PLINK function --*option-homozyg* to identify runs of homozygosity (ROH) using the following parameters: we considered at least 100 SNPs for ROH, with a total length ≥ 100 kilobases and at least one SNP per 50 kb on average; we set a scanning window to contain 100 SNPs, allowed 1 heterozygous call and 5 missing calls per scanning window. We used three component Gaussian mixture model from the Mclust package (v.6.1) in R [106] following Pemberton et al., [57] to classify the ROHs into short, intermediate and long sizes.

Demographic history. Utilizing the phased WGS data, we estimated effective population sizes at different time points within the Saudi sub-clusters using RELATE v1.1.9 [141]. We used the RelateFileFormats in the Relate package to convert files from VCF format into haps/sample file format. For ancestral allele flipping, we provided RELATE with the human ancestor sequences release 107. We computed the genealogical trees using the parameters -m 1.25e-8 -N 30,000 and subsequently used the EstimatePopulationSize.sh script provided with the Relate package to estimate the effective population sizes.

Evaluating imputation accuracy of Saudi genotypes

We evaluated the impact of Saudi's population demographics on the imputation accuracy of its haplotypes, using the Trans-Omics for Precision Medicine (TOPMed) panel [61]. We compared the imputation accuracy between the Saudi and the Europeans in UK10K dataset [62]. For this comparison, we first selected the variants that are present in both dataset and then subsetted the Europeans to 3,252 individuals in order to match the same number of individuals as our Saudi data.

Enrichment of functionally deleterious alleles

We compared the allelic architecture between Saudi Arabian and the gnomAD v4 [64,65] African/African American (gnomAD-AFR), non-Finnish European (gnomAD-EUR) and Middle Eastern (gnomAD-MID) populations. To check for potential enrichment or purging of deleterious

alleles in the Saudi, we computed the ratio of the proportional site frequency spectra for the deleterious alleles in Saudi to gnomAD-AFR or gnomAD-EUR, and contrasted it to the same ratio based on neutral or benign alleles. Utilizing the gnomAD exomes, which have a larger number of Middle Easterners compared to the genomes, we also made comparisons between the Middle Easterns and gnomAD-AFR and gnomAD-EUR. Significance differences in the ratios between variants functional classes were tested through bootstrapping.

For every comparison between populations or subpopulations, we used Hypergeometric (v 3.6.2) distribution in R [106] to downsample both populations to equal sample sizes. All exome comparisons were downsampled to gnomAD-MID sample size. To account for technical differences in data generation of WGS call sets between gnomAD and Saudi data, we used the proportions of variants from the normalized allele frequency spectra rather than number of variants to compare the ratio between the Saudi and the gnomAD populations at a given allele count or frequency bin. However, when comparisons were made between two gnomAD populations or between two Saudi subpopulations, the actual number of variants were used.

DATA AVAILABILITY

In compliance with Saudi privacy legislation and the protection of human subject confidentiality, the sharing of raw genotyping and clinical data is restricted. Access to this data requires prior approval from the Saudi National Bioethics Committee. The Saudi variants discovered through WGS and their estimated allele frequencies are deposited in the Figshare repository and can be accessed through this link: https://doi.org/10.6084/m9.figshare.28059686.v1.

ACKNOWLEDGMENTS

We are very grateful to the study participants who donated the samples used in the study. We thank Michael Campbell and Arun Durvasula for providing feedback on the drafted manuscript,

and Rui Leite Portela Martiniano for providing us with the Bahrain ancient DNA data. This work was supported by National Institute of General Medical Sciences (NIGMS) of the National Institute of Health under award number R35GM142783 (to C.W.K.C.). Computation for this work was supported by University of Southern California's Center for Advanced Research Computing (https://www.carc.usc.edu).

AUTHOR CONTRIBUTIONS

D.K.M. and C. W. K. C. conceived and designed the study. C. W. K. C., S.M., and M.A. acquired funding for the data generation and analysis in this study. M.A. performed sample acquisition and data generation. D.K.M. and M.P.W. analyzed the data. C.D.H. provided analysis tools. D.K.M., M.P.W., C.D.H. and C.W.K.C. interpreted the results. D.K.M., M.P.W. and C.W.K.C. wrote the manuscript with input from all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Word document: Supplementary Figures S1 – 17

Word document: Supplementary Tables S1, 2, 3 and 18

Excel spreadsheet: Supplementary Tables S4 – 17

REFERENCES

1. Armitage, S.J., Jasim, S.A., Marks, A.E., Parker, A.G., Usik, V.I., and Uerpmann H-P (2011). The Southern Route“Out of Africa”:Evidence for an Early Expansionof Modern Humans into Arabia. Science (1979) *331*, 453–456. https://doi.org/10.1594/PANGAEA.755114.

2. Henn, B.M., Cavalli-Sforza, L.L., and Feldman, M.W. (2012). The great human expansion. Preprint, https://doi.org/10.1073/pnas.1212380109 https://doi.org/10.1073/pnas.1212380109.

3. Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L., et al. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome Res *26*, 151–162. https://doi.org/10.1101/gr.191478.115.

4. Groucutt, H.S., Grün, R., Zalmout, I.A.S., Drake, N.A., Armitage, S.J., Candy, I., Clark-Wilson, R., Louys, J., Breeze, P.S., Duval, M., et al. (2018). Homo sapiens in Arabia by 85,000 years ago. Nat Ecol Evol *2*, 800–809. https://doi.org/10.1038/s41559-018-0518-2.

5. Fernandes, V., Alshamali, F., Alves, M., Costa, M.D., Pereira, J.B., Silva, N.M., Cherni, L., Harich, N., Cerny, V., Soares, P., et al. (2012). The Arabian cradle: Mitochondrial relicts of the first steps along the Southern route out of Africa. Am J Hum Genet *90*, 347–355. https://doi.org/10.1016/j.ajhg.2011.12.010.

6. Almarri, M.A., Haber, M., Lootah, R.A., Hallast, P., Al Turki, S., Martin, H.C., Xue, Y., and Tyler-Smith, C. (2021). The genomic history of the Middle East. Cell *184*, 4612-4625.e14. https://doi.org/10.1016/j.cell.2021.07.013.

7. Ferreira, J.C., Alshamali, F., Montinaro, F., Cavadas, B., Torroni, A., Pereira, L., Raveane, A., and Fernandes, V. (2021). Projecting Ancient Ancestry in Modern-Day Arabians and Iranians: A Key Role of the Past Exposed Arabo-Persian Gulf on Human Migrations. Genome Biol Evol *13*. https://doi.org/10.1093/gbe/evab194.

8. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature *536*, 419–424. https://doi.org/10.1038/nature19310.

9. Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The Date of Interbreeding between Neandertals and Modern Humans. PLoS Genet *8*. https://doi.org/10.1371/journal.pgen.1002947.

10. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413. https://doi.org/10.1038/nature13673.

11. Vallini, L., Zampieri, C., Shoaee, M.J., Bortolini, E., Marciani, G., Aneli, S., Pievani, T., Benazzi, S., Barausse, A., Mezzavilla, M., et al. (2024). The Persian plateau served as hub for Homo sapiens

after the main out of Africa dispersal. Nat Commun *15*. https://doi.org/10.1038/s41467-024-46161-7.

12. Martiniano, R., Haber, M., Almarri, M.A., Mattiangeli, V., Kuijpers, M.C.M., Chamel, B., Breslin, E.M., Littleton, J., Almahari, S., Aloraifi, F., et al. (2024). Ancient genomes illuminate Eastern Arabian population history and adaptation against malaria. Cell Genomics *4*. https://doi.org/10.1016/j.xgen.2024.100507.

13. Fernandes, V., Brucato, N., Ferreira, J.C., Pedro, N., Cavadas, B., Ricaut, F.X., Alshamali, F., and 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*, 575–586. https://doi.org/10.1093/molbev/msz005.

14. Khayat, A.M., Alshareef, B.G., Alharbi, S.F., AlZahrani, M.M., Alshangity, B.A., and Tashkandi, N.F. (2024). Consanguineous Marriage and Its Association With Genetic Disorders in Saudi Arabia: A Review. Cureus. https://doi.org/10.7759/cureus.53888.

15. Tadmouri, G.O., Nair, P., Obeid, T., Al Ali, M.T., Al Khaja, N., and Hamamy, H.A. (2009). Consanguinity and reproductive health among Arabs. Reprod Health *6*. https://doi.org/10.1186/1742-4755-6-17.

16. El-Hazmi, M.A.F., Al-Swailem, A.R., Warsy, A.S., Al-Swailem, A.M., Sulaimani, R., Al-Meshari, A.A., El-Hazmi, F., and Arabia, S. (1995). Consanguinity among the Saudi Arabian population.

17. Ben Halim, N., Ben Alaya Bouafif, N., Romdhane, L., Kefi Ben Atig, R., Chouchane, I., Bouyacoub, Y., Arfa, I., Cherif, W., Nouira, S., Talmoudi, F., et al. (2013). Consanguinity, endogamy, and genetic disorders in Tunisia. J Community Genet *4*, 273–284. https://doi.org/10.1007/s12687-012-0128-7.

18. Bittles, A.H. (2008). A community genetics perspective on consanguineous marriage. Preprint, https://doi.org/10.1159/000133304 https://doi.org/10.1159/000133304.

19. Alkuraya, F.S. (2014). Genetics and genomic medicine in saudi arabia. Mol Genet Genomic Med *2*, 369–378. https://doi.org/10.1002/mgg3.97.

20. Mineta, K., Goto, K., Gojobori, T., and Alkuraya, F.S. (2021). Population structure of indigenous inhabitants of Arabia. PLoS Genet *17*, e1009210. https://doi.org/10.1371/JOURNAL.PGEN.1009210.

21. Temaj, G., Nuhii, N., and Sayer, J.A. (2022). The impact of consanguinity on human health and disease with an emphasis on rare diseases. Journal of Rare Diseases *1*. https://doi.org/10.1007/s44162-022-00004-5.

22. Aleissa, M., Aloraini, T., Alsubaie, L.F., Hassoun, M., Abdulrahman, G., Swaid, A., Al Eyaid, W., Al Mutairi, F., Ababneh, F., Alfadhel, M., et al. (2022). Common disease-associated gene variants in a Saudi Arabian population. Ann Saudi Med *42*, 29–35. https://doi.org/10.5144/0256-4947.2022.29.

23. Delatycki, M.B., Alkuraya, F., Archibald, A., Castellani, C., Cornel, M., Grody, W.W., Henneman, L., Ioannides, A.S., Kirk, E., Laing, N., et al. (2020). International perspectives on the implementation

of reproductive carrier screening. Preprint at John Wiley and Sons Ltd, https://doi.org/10.1002/pd.5611 https://doi.org/10.1002/pd.5611.

24. Hedrick, P.W., and Garcia-Dorado, A. (2016). Understanding Inbreeding Depression, Purging, and Genetic Rescue. Preprint at Elsevier Ltd, https://doi.org/10.1016/j.tree.2016.09.005 https://doi.org/10.1016/j.tree.2016.09.005.

25. Sahoo, S.A., Zaidi, A.A., Anagol, S., and Mathieson, I. (2021). Long Runs of Homozygosity Are Correlated with Marriage Preferences across Global Population Samples. Hum Biol *93*, 201–216. https://doi.org/10.1353/hub.2021.0011.

26. Scott, E.M., Halees, A., Itan, Y., Spencer, E.G., He, Y., Azab, M.A., Gabriel, S.B., Belkadi, A., Boisson, B., Abel, L., et al. (2016). Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. Preprint at Nature Research, https://doi.org/10.1038/ng.3592 https://doi.org/10.1038/ng.3592.

27. Alsalem, A.B., Halees, A.S., Anazi, S., Alshamekh, S., and Alkuraya, F.S. (2013). Autozygome Sequencing Expands the Horizon of Human Knockout Research and Provides Novel Insights into Human Phenotypic Variation. PLoS Genet *9*. https://doi.org/10.1371/journal.pgen.1004030.

28. Overall, A., Ahmad, M., and Nichols, R.A. (2002). The effect of reproductive compensation on recessive disorders within consanguineous human populations. Heredity (Edinb) *88*, 474–479. https://doi.org/10.1038/sj/hdy/6800090.

29. Ober, C., Hyslop, T., and Hauck, W.W. (1999). Inbreeding Effects on Fertility in Humans: Evidence for Reproductive Compensation.

30. Lohmueller, K.E. (2014). The distribution of deleterious genetic variation in human populations. Preprint at Elsevier Ltd, https://doi.org/10.1016/j.gde.2014.09.005 https://doi.org/10.1016/j.gde.2014.09.005.

31. Castellano, S., Parra, G., Sánchez-Quinto, F.A., Racimo, F., Kuhlwilm, M., Kircher, M., Sawyer, S., Fu, Q., Heinze, A., Nickel, B., et al. (2014). Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A *111*, 6666–6671. https://doi.org/10.1073/pnas.1405138111.

32. Simons, Y.B., and Sella, G. (2016). The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Preprint at Elsevier Ltd, https://doi.org/10.1016/j.gde.2016.09.006 https://doi.org/10.1016/j.gde.2016.09.006.

33. Laurent, R., Gineau, L., Utge, J., Lafosse, S., Phoeung, C.L., Hegay, T., Olaso, R., Boland, A., Deleuze, J.F., Toupance, B., et al. (2024). Measuring the Efficiency of Purging by non-random Mating in Human Populations. Mol Biol Evol *41*. https://doi.org/10.1093/molbev/msae094.

34. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science (1979) *367*. https://doi.org/10.1126/science.aay5012.

35. Elliott, K.S., Haber, M., Daggag, H., Busby, G.B., Sarwar, R., Kennet, D., Petraglia, M., Petherbridge, L.J., Yavari, P., Heard-Bey, F.U., et al. (2022). Fine-Scale Genetic Structure in the United Arab

Emirates Reflects Endogamous and Consanguineous Culture, Population History, and Geography. Mol Biol Evol *39*. https://doi.org/10.1093/molbev/msac039.

36. Hunwick, J.O. (2006). Arab views of black Africans and slavery. In West Africa, Islam, and the Arab world, pp. 75–90.

37. Charati, H., and Ori, R.J. (2021). Patterns of Genetic Structure and Evidence of Gene Flow between Arabian Peninsula and European Populations. Am J Biomed Sci Res *12*, 285–291. https://doi.org/10.34297/ajbsr.2021.12.001759.

38. Marcus, J.H., Posth, C., Ringbauer, H., Lai, L., Skeates, R., Sidore, C., Beckett, J., Furtwängler, A., Olivieri, A., Chiang, C.W.K., et al. (2020). Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia. Nat Commun *11*. https://doi.org/10.1038/s41467-020-14523-6.

39. Chiang, C.W.K., Marcus, J.H., Sidore, C., Biddanda, A., Al-Asadi, H., Zoledziewska, M., Pitzalis, M., Busonero, F., Maschio, A., Pistis, G., et al. (2018). Genomic history of the Sardinian population. Nat Genet *50*, 1426–1434. https://doi.org/10.1038/s41588-018-0215-8.

40. Fernandes, D.M., Mittnik, A., Olalde, I., Lazaridis, I., Cheronet, O., Rohland, N., Mallick, S., Bernardos, R., Broomandkhoshbacht, N., Carlsson, J., et al. (2020). The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. Nat Ecol Evol *4*, 334–345. https://doi.org/10.1038/s41559-020-1102-0.

41. Razali, R.M., Rodriguez-Flores, J., Ghorbani, M., Naeem, H., Aamer, W., Aliyev, E., Jubran, A., Ismail, S.I., Al-Muftah, W., Badji, R., et al. (2021). Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. Nat Commun *12*. https://doi.org/10.1038/s41467-021-25287-y.

42. Lebling, R.W. (2009). “The Saracens of St. Tropez.” Saudi Aramco World.

43. Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., and Reich, D. (2024). The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. Sci Data *11*. https://doi.org/10.1038/s41597-024-03031-7.

44. Pagani, L., and Pagani, L.& C.I. (2019). What is Africa? A human perspective in Modern human origins and dispersal. In Morden human origins and dispersal, H. Katerina and J. Gerhard, eds. (Kerns Verlag), pp. 15–24.

45. Yang, M.A., and Fu, Q. (2018). Insights into Modern Human Prehistory Using Ancient Genomes. Preprint at Elsevier Ltd, https://doi.org/10.1016/j.tig.2017.11.008 https://doi.org/10.1016/j.tig.2017.11.008.

46. Chintalapati, M., Patterson, N., and Moorjani, P. (2022). The spatiotemporal patterns of major human admixture events during the European Holocene. Elife *11*. https://doi.org/10.7554/ELIFE.77625.

47. Antonio, M.L., Weiß, C.L., Gao, Z., Sawyer, S., Oberreiter, V., Moots, H.M., Spence, J.P., Cheronet, O., Zagorc, B., Praxmarer, E., et al. (2024). Stable population structure in Europe since the Iron Age, despite high mobility. Elife *13*. https://doi.org/10.7554/ELIFE.79714.

48. Rodríguez-Varela, R., Günther, T., Krzewińska, M., Storå, J., Gillingwater, T.H., MacCallum, M., Arsuaga, J.L., Dobney, K., Valdiosera, C., Jakobsson, M., et al. (2017). Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. Curr Biol *27*, 3396-3402.e5. https://doi.org/10.1016/J.CUB.2017.09.059.

49. Van De Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E.H., Abdeljalil, M., et al. (2018). Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. Science (1979), 548–552.

50. Moots, H.M., Antonio, M., Sawyer, S., Spence, J.P., Oberreiter, V., Weiß, C.L., Lucci, M., Cherifi, Y.M.S., La Pastina, F., Genchi, F., et al. (2023). A genetic history of continuity and mobility in the Iron Age central Mediterranean. Nature Ecology & Evolution 2023 7:9 *7*, 1515–1524. https://doi.org/10.1038/s41559-023-02143-4.

51. Schuenemann, V.J., Peltzer, A., Welte, B., Van Pelt, W.P., Molak, M., Wang, C.C., Furtwängler, A., Urban, C., Reiter, E., Nieselt, K., et al. (2017). Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. Nat Commun *8*. https://doi.org/10.1038/ncomms15694.

52. Prendergast, M.E., Lipson, M., Sawchuk, E.A., Olalde, I., Ogola, C.A., Rohland, N., Sirak, K.A., Adamski, N., Bernardos, R., Broomandkhoshbacht, N., et al. (2019). Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. Science (1979) *364*.

53. Sirak, K.A., Fernandes, D.M., Lipson, M., Mallick, S., Mah, M., Olalde, I., Ringbauer, H., Rohland, N., Hadden, C.S., Harney, É., et al. (2021). Social stratification without genetic differentiation at the site of Kulubnarti in Christian Period Nubia. Nature Communications 2021 12:1 *12*, 1–14. https://doi.org/10.1038/s41467-021-27356-8.

54. Lipson, M., Sawchuk, E.A., Thompson, J.C., Oppenheimer, J., Tryon, C.A., Ranhorn, K.L., de Luna, K.M., Sirak, K.A., Olalde, I., Ambrose, S.H., et al. (2022). Ancient DNA and deep population structure in sub-Saharan African foragers. Nature 2022 603:7900 *603*, 290–296. https://doi.org/10.1038/s41586-022-04430-9.

55. Schlebusch, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Vicente, M., Steyn, M., Soodyall, H., et al. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science (1979) *358*, 652–655.

56. Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: Windows into population history and trait architecture. Preprint at Nature Publishing Group, https://doi.org/10.1038/nrg.2017.109 https://doi.org/10.1038/nrg.2017.109.

57. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet *91*, 275–292. https://doi.org/10.1016/j.ajhg.2012.06.014.

58. Thompson, E.A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. Preprint, https://doi.org/10.1534/genetics.112.148825 https://doi.org/10.1534/genetics.112.148825.

59. Ceballos, F.C., Gürün, K., Altınışık, N.E., Gemici, H.C., Karamurat, C., Koptekin, D., Vural, K.B., Mapelli, I., Sağlıcan, E., Sürer, E., et al. (2021). Human inbreeding has decreased in time through the Holocene. Current Biology *31*, 3925-3934.e8. https://doi.org/10.1016/j.cub.2021.06.027.

60. Petraglia, M.D., Groucutt, H.S., Guagnin, M., Breeze, P.S., and Boivin, N. (2020). Human responses to climate and ecosystem change in ancient Arabia. Proceedings of the National Academy of Sciences *117*, 8263–8270. https://doi.org/10.1073/pnas.1920211117/-/DCSupplemental.

61. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.

62. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–89. https://doi.org/10.1038/nature14962.

63. Cahoon, J.L., Rui, X., Tang, E., Simons, C., Langie, J., Chen, M., Lo, Y.-C., and Chiang, C.W.K. (2024). Imputation Accuracy Across Global Human Populations. Am J Hum Genet *111*, P979-989. https://doi.org/10.1101/2023.05.22.541241.

64. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2023). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92–100. https://doi.org/10.1038/s41586-023-06045-0.

65. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

66. Ljungdahl, A., Kohani, S., Page, N.F., Wells, E.S., Wigdor, E.M., Dong, S., and Sanders, S.J. (2023). AlphaMissense is better correlated with functional assays of missense impact than earlier prediction algorithms. bioRxiv , 562294. https://doi.org/10.1101/2023.10.24.562294.

67. Benegas, G.I., Singh Batra, S.I., Song, Y.S., and Edited by Kathryn Roeder, I. (2023). BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS DNA language models are powerful predictors of genome-wide variant effects. *120*. https://doi.org/10.1073/pnas.

68. Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Ryan, D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R., et al. (2008). Proportionally More Deleterious Genetic Variation In European than in African Populations. Nature *451*, 994–997. https://doi.org/10.1038/nature06611.Proportionally.

69. Henn, B.M., Botigué, L.R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder, M.P., et al. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proceedings of the National Academy of Sciences *113*, E440–E449. https://doi.org/10.1073/pnas.1510805112.

70. Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. Nat Genet *46*, 220–224. https://doi.org/10.1038/ng.2896.

71. Do, R., Balick, D., Li, H., Adzhubei, I., Sunyaev, S., and Reich, D. (2015). No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet *47*, 126–131. https://doi.org/10.1038/ng.3186.

72. Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., et al. (2014). Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. PLoS Genet *10*. https://doi.org/10.1371/journal.pgen.1004494.

73. Pedersen, C.E.T., Lohmueller, K.E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H.R., Moltke, I., and Albrechtsen, A. (2017). The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: Insights from the Greenlandic Inuit. Genetics *205*, 787–801. https://doi.org/10.1534/genetics.116.193821.

74. Locke, A.E., Steinberg, K.M., Chiang, C.W.K., Service, S.K., Havulinna, A.S., Stell, L., Pirinen, M., Abel, H.J., Chiang, C.C., Fulton, R.S., et al. (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. Nature *572*, 323–328. https://doi.org/10.1038/s41586-019-1457-z.

75. Subramanian, S. (2016). Europeans have a higher proportion of high-frequency deleterious variants than Africans. Hum Genet *135*, 1–7. https://doi.org/10.1007/s00439-015-1604-z.

76. Eaaswarkhanth, M., Pathak, A.K., Ongaro, L., Montinaro, F., Prashantha Hebbar, •, Osama Alsmadi, •, Metspalu, M., Al-Mulla, F., Thangavel, •, and Thanaraj, A. (2021). Unraveling a fine-scale high genetic heterogeneity and recent continental connections of an Arabian Peninsula population. European Journal of Human Genetics *30*, 307–319. https://doi.org/10.1038/s41431-021-00861-6.

77. Khubrani, Y.M., Wetton, J.H., and Jobling, M.A. (2018). Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. Forensic Sci Int Genet *33*, 98–105. https://doi.org/10.1016/j.fsigen.2017.11.015.

78. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the Population Genetic History of the Caribbean. PLoS Genet *9*. https://doi.org/10.1371/journal.pgen.1003925.

79. Browning, S.R., Grinde, K., Plantinga, A., Gogarten, S.M., Stilp, A.M., Kaplan, R.C., Avilés-Santa, M.L., Browning, B.L., and Laurie, C.C. (2016). Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic community health study/study of Latinos (HCHS/SOL). G3: Genes, Genomes, Genetics *6*, 1525–1534. https://doi.org/10.1534/g3.116.028779.

80. El-Mouzan, M.I., Al-Salloum, A.A., Al-Herbish, A.S., Qurachi, M.M., and Al-Omar, A.A. (2007). Regional variations in the prevalence of consanguinity in Saudi Arabia. Saudi Med J *28*.

81. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science (1979) *343*, 747–751. https://doi.org/10.1126/science.1243518.

82. Miran, J. (2022). Red Sea Slave Trade (Oxford University Press) https://doi.org/10.1093/acrefore/9780190277734.013.868.

83. Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.C., Gbeha, E., Hamdan, F.F., Girard, S., et al. (2013). Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. PLoS Genet *9*. https://doi.org/10.1371/journal.pgen.1003815.

84. Warsy, A.S., Al-Jaser, M.H., Albdass, A., Al-Daihan, S., and Alanazi, M. (2014). Is consanguinity prevalence decreasing in Saudis?: A study in two generations. Afr Health Sci *14*, 314–321. https://doi.org/10.4314/ahs.v14i2.5.

85. Albanghali, M.A. (2023). Prevalence of Consanguineous Marriage among Saudi Citizens of Albaha, a Cross-Sectional Study. Int J Environ Res Public Health *20*. https://doi.org/10.3390/ijerph20043767.

86. Al-Gazali, L., Hamamy, H., and Al-Arrayad, S. (2006). Genetic disorders in the Arab world.

87. Saffi, M., and Howard, N. (2015). Exploring the Effectiveness of Mandatory Premarital Screening and Genetic Counselling Programmes for β-Thalassaemia in the Middle East: A Scoping Review. Preprint at S. Karger AG, https://doi.org/10.1159/000430837 https://doi.org/10.1159/000430837.

88. Tadmouri, G.O., Nair, P., Obeid, T., Al Ali, M.T., Al Khaja, N., and Hamamy, H.A. (2009). Consanguinity and reproductive health among Arabs. Reprod Health *6*. https://doi.org/10.1186/1742-4755-6-17.

89. Elfatih, A., Saad, C., Ismail, S., Al-Muftah, W., Badji, R., Darwish, D., Fadl, T., Yasin, H., Ennaifar, M., Abdel-latif, R., et al. (2024). Analysis of 14,392 whole genomes reveals 3.5% of Qataris carry medically actionable variants. European Journal of Human Genetics. https://doi.org/10.1038/s41431-024-01656-1.

90. Mbarek, H., Gandhi, G.D., Selvaraj, S., Al-Muftah, W., Badji, R., Al-Sarraj, Y., Saad, C., Darwish, D., Alvi, M., Fadl, T., et al. (2022). Qatar genome: Insights on genomics from the Middle East. Hum Mutat. https://doi.org/10.1002/HUMU.24336.

91. Thareja, G., Al-Sarraj, Y., Belkadi, A., Almotawa, M., Ismail, S., Al-Muftah, W., Badji, R., Mbarek, H., Darwish, D., Fadl, T., et al. (2021). Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. Nature Communications 2021 12:1 *12*, 1–10. https://doi.org/10.1038/s41467-021-21381-3.

92. Kousathanas, A., Pairo-Castineira, E., Rawlik, K., Stuckey, A., Odhams, C.A., Walker, S., Russell, C.D., Malinauskas, T., Wu, Y., Millar, J., et al. (2022). Whole-genome sequencing reveals host factors underlying critical COVID-19. Nature *607*, 97–103. https://doi.org/10.1038/s41586-022-04576-6.

93. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559–575. https://doi.org/10.1086/519795.

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

95. Browning, B.L., Tian, X., Zhou, Y., and Browning, S.R. (2021). Fast two-stage phasing of large-scale sequence data. https://doi.org/10.1016/j.ajhg.2021.08.005.

96. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

97. Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T.L., Huang, J., Vujkovic, M., Damrauer, S.M., Pyarajan, S., Gaziano, J.M., et al. (2019). Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. Am J Hum Genet *105*, 763–772. https://doi.org/10.1016/j.ajhg.2019.08.012.

98. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

99. Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

100. Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., et al. (2017). Data Descriptor: Whole genome characterization of sequence diversity of 15,220 Icelanders. Sci Data *4*. https://doi.org/10.1038/sdata.2017.115.

101. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. https://doi.org/10.1002/0471250953.bi1110s43.

102. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet *48*, 1443–1448. https://doi.org/10.1038/ng.3679.

103. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070. https://doi.org/10.1093/bioinformatics/btq330.

104. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulyte, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (1979) *381*. https://doi.org/10.1126/science.adg7492.

105. Mcinnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection Software • Review • Repository • Archive. https://doi.org/10.21105/joss.00861.

106. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

107. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.

108. Batool, F., and Hennig, C. (2019). Clustering with the Average Silhouette Width.

109. Alexander, D.H., and Novembre, J. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. 1655–1664. https://doi.org/10.1101/gr.094052.109.vidual.

110. Maier, R., and Patterson, N. (2024). admixtools: Inferring demographic history from genetic data. R package version 2.0.4. Preprint.

111. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

112. Maier, R., Flegontov, P., Flegontova, O., Işıldak, U., Changmai, P., and Reich, D. (2023). On the limits of fitting complex models of population history to f-statistics. Elife *12*. https://doi.org/10.7554/eLife.85492.

113. Gallego Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science (1979) *350*, 820–822.

114. Wang, K., Goldstein, S., Bleasdale, M., Clist, B., Clist, B., Bostoen, K., Bakwa-Lufu, P., Buck, L.T., Buck, L.T., Crowther, A., et al. (2020). Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. Sci Adv *6*, 183–195.

115. Skoglund, P., Thompson, J.C., Prendergast, M.E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., et al. (2017). Reconstructing Prehistoric African Population Structure. Cell *171*, 59-71.e21. https://doi.org/10.1016/J.CELL.2017.08.049.

116. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature 2008 451:7181 *451*, 998–1003. https://doi.org/10.1038/nature06742.

117. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships

1275 inferred from genome-wide patterns of variation. Science *319*, 1100–1104.
1276 https://doi.org/10.1126/SCIENCE.1153717.

1277 118. Narasimhan, V.M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., Lazaridis, I.,
1278 Nakatsuka, N., Olalde, I., Lipson, M., et al. (2019). The formation of human populations in South
1279 and Central Asia. Science (1979) *365*.

1280 119. Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin,
1281 R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., et al. (2015). Upper Palaeolithic genomes
1282 reveal deep roots of modern Eurasians. Nature Communications 2015 6:1 *6*, 1–8.
1283 https://doi.org/10.1038/ncomms9912.

1284 120. Skoglund, P., Mallick, S., Bortolini, M.C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M.L., Salzano,
1285 F.M., Patterson, N., and Reich, D. (2015). Genetic evidence for two founding populations of the
1286 Americas. Nature 2015 525:7567 *525*, 104–108. https://doi.org/10.1038/nature14895.

1287 121. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N.,
1288 Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from
1289 142 diverse populations. Nature *538*, 201–206. https://doi.org/10.1038/nature18964.

1290 122. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Aximu-Petri,
1291 A., Prüfer, K., De Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human
1292 from western Siberia. Nature 2014 514:7523 *514*, 445–449.
1293 https://doi.org/10.1038/nature13810.

1294 123. Prüfer, K., De Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L.,
1295 Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in
1296 Croatia. Science (1979) *358*, 655–658.

1297 124. Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S.,
1298 Stafford, T.W., Orlando, L., Metspalu, E., et al. (2013). Upper Palaeolithic Siberian genome reveals
1299 dual ancestry of Native Americans. Nature 2013 505:7481 *505*, 87–91.
1300 https://doi.org/10.1038/nature12736.

1301 125. Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W.,
1302 Meyer, M., Mittnik, A., et al. (2016). The genetic history of Ice Age Europe. Nature 2016 534:7606
1303 *534*, 200–205. https://doi.org/10.1038/nature17993.

1304 126. Olalde, I., Allentoft, M.E., Sánchez-Quinto, F., Santpere, G., Chiang, C.W.K., DeGiorgio, M., Prado-
1305 Martinez, J., Rodríguez, J.A., Rasmussen, S., Quilez, J., et al. (2014). Derived immune and ancestral
1306 pigmentation alleles in a 7,000-year-old Mesolithic European. Nature 2014 507:7491 *507*, 225–
1307 228. https://doi.org/10.1038/nature12960.

1308 127. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E.,
1309 Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230
1310 ancient Eurasians. Nature *528*, 499–503. https://doi.org/10.1038/NATURE16152.

1311 128. Feldman, M., Fernández-Domínguez, E., Reynolds, L., Baird, D., Pearson, J., Hershkovitz, I., May,
1312 H., Goring-Morris, N., Benz, M., Gresky, J., et al. (2019). Late Pleistocene human genome suggests

a local origin for the first farmers of central Anatolia. Nature Communications 2019 10:1 *10*, 1–10. https://doi.org/10.1038/s41467-019-09209-7.

129. Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., Akyüz, U., Andreeva, D., Andrijašević, G., Antonović, D., et al. (2022). The genetic history of the Southern Arc: A bridge between West Asia and Europe. Science (1979) *377*.

130. Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., Akyüz, U., Andreeva, D., Andrijašević, G., Antonović, D., et al. (2022). Ancient DNA from Mesopotamia suggests distinct Pre-Pottery and Pottery Neolithic migrations into Anatolia. Science (1979) *377*, 982–987.

131. Yaka, R., Mapelli, I., Kaptan, D., Doğu, A., Chyleński, M., Erdal, Ö.D., Koptekin, D., Vural, K.B., Bayliss, A., Mazzucato, C., et al. (2021). Variable kinship patterns in Neolithic Anatolia revealed by ancient genomes. Curr Biol *31*, 2455-2468.e18. https://doi.org/10.1016/J.CUB.2021.03.050.

132. Harney, É., May, H., Shalem, D., Rohland, N., Mallick, S., Lazaridis, I., Sarig, R., Stewardson, K., Nordenfelt, S., Patterson, N., et al. (2018). Publisher Correction: Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. Nature Communications 2018 9:1 *9*, 1–1. https://doi.org/10.1038/s41467-018-06484-8.

133. Wang, C.C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., Cheronet, O., Ferry, M., Harney, E., Keating, D., et al. (2019). Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nature Communications 2019 10:1 *10*, 1–13. https://doi.org/10.1038/s41467-018-08220-8.

134. Agranat-Tamir, L., Waldman, S., Martin, M.A.S., Gokhman, D., Mishol, N., Eshel, T., Cheronet, O., Rohland, N., Mallick, S., Adamski, N., et al. (2020). The Genomic History of the Bronze Age Southern Levant. Cell *181*, 1146-1157.e11. https://doi.org/10.1016/j.cell.2020.04.024.

135. Haber, M., Doumet-Serhal, C., Scheib, C., Xue, Y., Danecek, P., Mezzavilla, M., Youhanna, S., Martiniano, R., Prado-Martinez, J., Szpak, M., et al. (2017). Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences. Am J Hum Genet *101*, 274–282. https://doi.org/10.1016/J.AJHG.2017.06.013.

136. Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. Nature *555*, 190–196. https://doi.org/10.1038/NATURE25738.

137. Patterson, N., Isakov, M., Booth, T., Büster, L., Fischer, C.E., Olalde, I., Ringbauer, H., Akbari, A., Cheronet, O., Bleasdale, M., et al. (2021). Large-scale migration into Britain during the Middle to Late Bronze Age. Nature 2021 601:7894 *601*, 588–594. https://doi.org/10.1038/s41586-021-04287-4.

138. Feldman, M., Master, D.M., Bianco, R.A., Burri, M., Stockhammer, P.W., Mittnik, A., Aja, A.J., Jeong, C., and Krause, J. (2019). Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. Sci Adv *5*, 61–64.

139. Moorjani, P., Sankararaman, S., Fu, Q., Przeworski, M., Patterson, N., and Reich, D. (2016). A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proc Natl Acad Sci U S A *113*, 5652–5657. https://doi.org/10.1073/pnas.1514696113.

140. Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., et al. (2020). High-depth African genomes inform human migration and health. Nature *586*, 741–748. https://doi.org/10.1038/s41586-020-2859-7.

141. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. Nat Genet *51*, 1321–1329. https://doi.org/10.1038/s41588-019-0484-x.