RESEARCH ARTICLE

Population structure of indigenous inhabitants of Arabia

Katsuhiko Mineta[1]☯, **Kosuke Goto**[1]☯, **Takashi Gojobori**[1]*, **Fowzan S. Alkuraya**[2,3,4]*

1 Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, **2** Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia, **3** Department of Anatomy and Cell Biology, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia, **4** Saudi Human Genome Program, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

☯ These authors contributed equally to this work.
* takashi.gojobori@kaust.edu.sa (TG); falkuraya@kfshrc.edu.sa (FSA)

OPEN ACCESS

Citation: Mineta K, Goto K, Gojobori T, Alkuraya FS (2021) Population structure of indigenous inhabitants of Arabia. PLoS Genet 17(1): e1009210. https://doi.org/10.1371/journal.pgen.1009210

Editor: Youssef Idaghdour, NCSU, UNITED STATES

Received: February 6, 2020

Accepted: October 16, 2020

Published: January 11, 2021

Data Availability Statement: Tribal genotypes are available as supporting information (S1 Data). Individual data cannot be shared publicly because of the consent restrictions and is available from the KFSH&RC Institutional Ethics Committee (contact via Dr. Abeer Omer: aomer@kfshrc.edu.sa) for researchers who meet the criteria for access to confidential data.

Funding: We acknowledge the support of the Saudi Human Genome Project, King Abdulaziz City for Science and Technology to FSA and King Abdullah University of Science and Technology (BAS/1/

Abstract

Modern day Saudi Arabia occupies the majority of historical Arabia, which may have contributed to ancient waves of migration out of Africa. This ancient history has left a lasting imprint in the genetics of the region, including the diverse set of tribes that call Saudi Arabia their home. How these tribes relate to each other and to the world's major populations remains an unanswered question. In an attempt to improve our understanding of the population structure of Saudi Arabia, we conducted genomic profiling of 957 unrelated individuals who self-identify with 28 large tribes in Saudi Arabia. Consistent with the tradition of intra-tribal unions, the subjects showed strong clustering along tribal lines with the distance between clusters correlating with their geographical proximities in Arabia. However, these individuals form a unique cluster when compared to the world's major populations. The ancient origin of these tribal affiliations is supported by analyses that revealed little evidence of ancestral origin from within the 28 tribes. Our results disclose a granular map of population structure and have important implications for future genetic studies into Mendelian and common diseases in the region.

Author summary

Arabian Peninsula has been home to ancient civilizations and elucidating its population structure can add to our understanding of human history. We conducted genomic profiling of 957 unrelated individuals who self-identify with 28 large tribes in Saudi Arabia. We found that genomic structure of Saudis is more clustered than previously thought, suggesting that tribal linages have been maintained for a long time in Arabian Peninsula in spite of admixture. The population structure revealed by our study has important implications for both Mendelian and complex genetics work in the study population.

1059/01/01) to TG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Early studies on human genome have revealed a marked degree of similarity between seemingly distinct human populations [1,2]. However, the small fraction of the human variome that shows a substantial difference between different ancestries has a significant impact on a wide range of medical applications of genetic and genomics. For example, in the area of complex disease genetics, it is important to compare samples with a genetically matched background in order to avoid false attribution of associations to population stratification [3]. Similarly, correctly identifying causal variants in Mendelian diseases is critically dependent on the knowledge of allele frequency in different populations. Indeed, studies have shown how gaps in this knowledge can lead to erroneous disease links [4,5]. Proper understanding of population structure also allows for increased representation of diverse populations in large genome-wide associations studies (GWAS) to accelerate the discovery of genetic risk loci for common diseases and the implementation of this knowledge in calculating polygenic risk scores that can be adjusted based on ancestry [6,7].

In addition to these practical implications, population diversity is key to retracing early human history. Dendrograms constructed based on genetic variation within human populations have painted a picture of human migration out of Africa that correlates well with geographic landmarks and historical accounts. Our ability to identify the ancestry of ancient human remains is greatly facilitated by detailed maps of the genetic structure of contemporary human populations as demonstrated by a number of high profile archeological finds [8,9]. Conversely, gaps in these maps caused by failure to represent certain ancestries can substantially reduce the accuracy with which ancient human remains are traced to specific ancestries. It is encouraging to see a rapid proliferation of population genetics projects that are gradually filling in the gaps created by the historically skewed representation of the world's population.

Arabian Peninsula has been home to ancient human civilizations, and human remains that date to > 85,000 years, and even much older artefacts from modern humans, have been recovered in archaeological sites in Arabia [10]. Indeed, accumulating genetic evidence suggests an exodus route out of Africa that may have involved the "Arabian cradle" [11]. Tribalism is a deeply rooted tradition in Arabia. The tribe, here, is a social human group, which is based on its cultural, anthropological, geographical disciplines. In Saudi Arabia, tribal affiliation can easily be determined by the common practice of using the tribe name as the surname, which is strictly inherited patrilineally. To enhance loyalty to the tribe, intra-tribal and, in the more restrictive form, intrafamilial (consanguineous) unions are encouraged [12]. We have analyzed in previous work the imprint of this endogamy on the genetic landscape of contemporary Saudi Arabians, and challenged the notion that long standing endogamy leads to "purging" of deleterious alleles [13,14]. However, the influence of tribal affiliation on the local population structure has not been well studied. In addition, large global efforts to characterize the human variome historically lacked adequate representation of natives of Arabia, potentially undermining ongoing disease studies in this region [15,16]. While there have been some attempts to investigate the population structure and diversity in the Arabian Peninsula from Kuwait [17] and Qatar [18,19], the bulk of the population in Arabia who reside in Saudi Arabia have yet to be studies. In this study, we aim at revealing the genetic structure of the Saudi Arabians by the tribal affiliations identified by their surnames. We present data on 957 unrelated individuals who represent 28 large tribes of Saudi Arabia. Our data reveal previously unrecognized influence of tribalism on genetic structure with important practical implications on the delivery of genomic medicine.

Results

A distinct profile of indigenous Arabian peninsula's population

The genotyping platform used in our analysis is incorporated with public resources of human variations such as HapMap, 1000 genomes and dbSNP. However, our analysis shows that nearly 5% of these SNPs were excluded after MAF filtering. This bias is not surprising since the Saudi population is largely lacking in the public SNP databases that were used to design this genotyping microarray. Consistent with this apparent difference in genetic makeup, when we compared the genotyping profile of our cohort to the more commonly studied continental populations, we observed that Saudi Arabians form a unique cluster (S1 Fig; S2–S12 Figs for details). Furthermore, the results of ADMIXTURE show that the major world populations (Africa, Europe, Asia, America and Oceania) and Saudi population are distinguished (K = 7) and tribes in Saudi are divided as K increased by the minimum error (K = 18) (Figs 1 and S13). We observed very little genetic contribution of these major world populations in our cohort

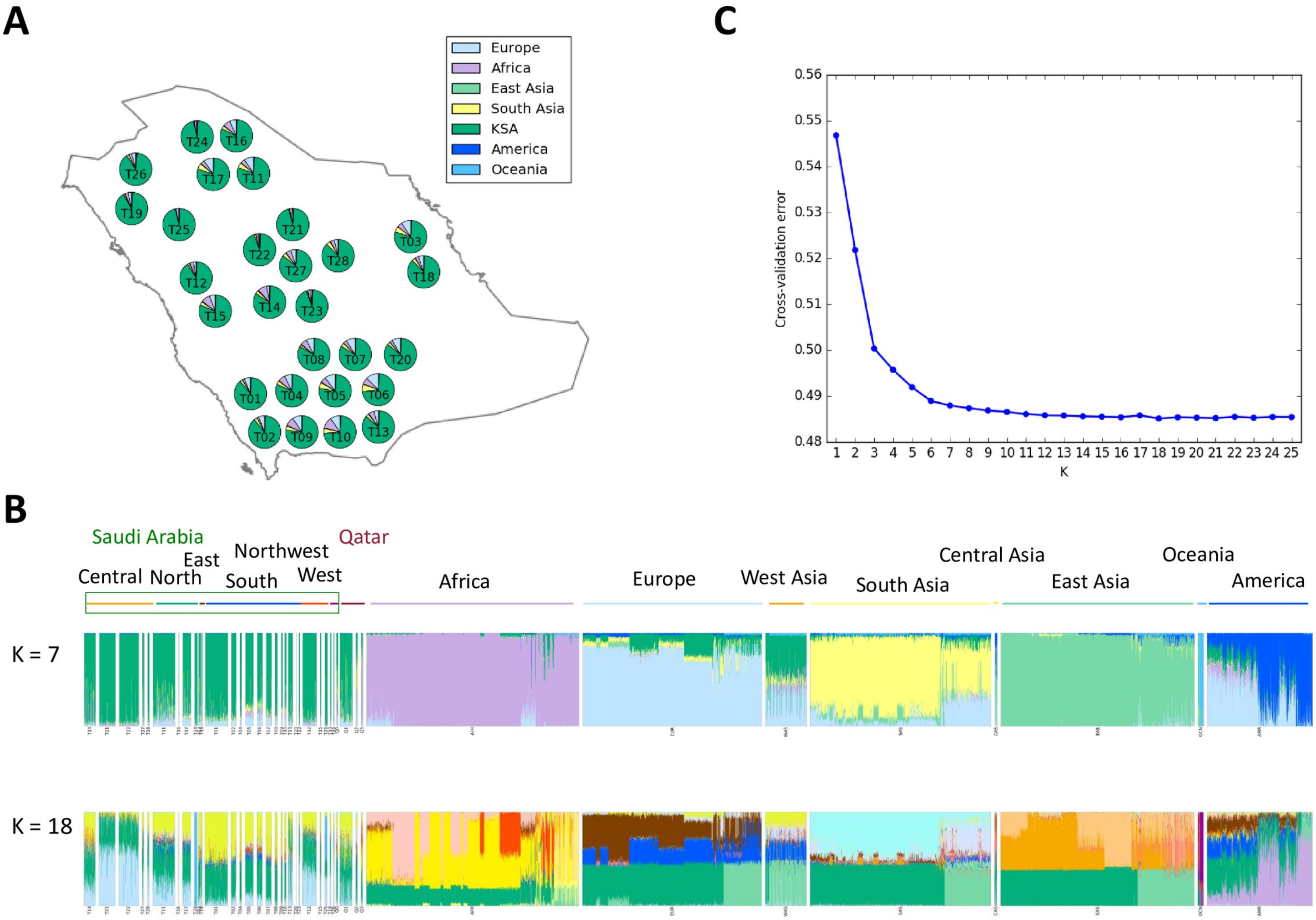


Fig 1. Saudi population structure. (A) Map of sampled tribes and admixture average proportions. Pie charts summarize per-tribe average proportions of Indigenous Arab samples at K = 7. Assumed ancestries are Arab, European, African, East Asian, South Asian, American and Oceanian. The map was made with Natural Earth (public domain). (B) Global ancestry proportions at K = 7 (top) and K = 18 (bottom) estimated with ADMIXTURE including African, European, Asian, American, Oceanian and Arab. (C) Cross-validation error for K runs from 1 to 25, K = 18 has the lowest cross-validation error.

https://doi.org/10.1371/journal.pgen.1009210.g001

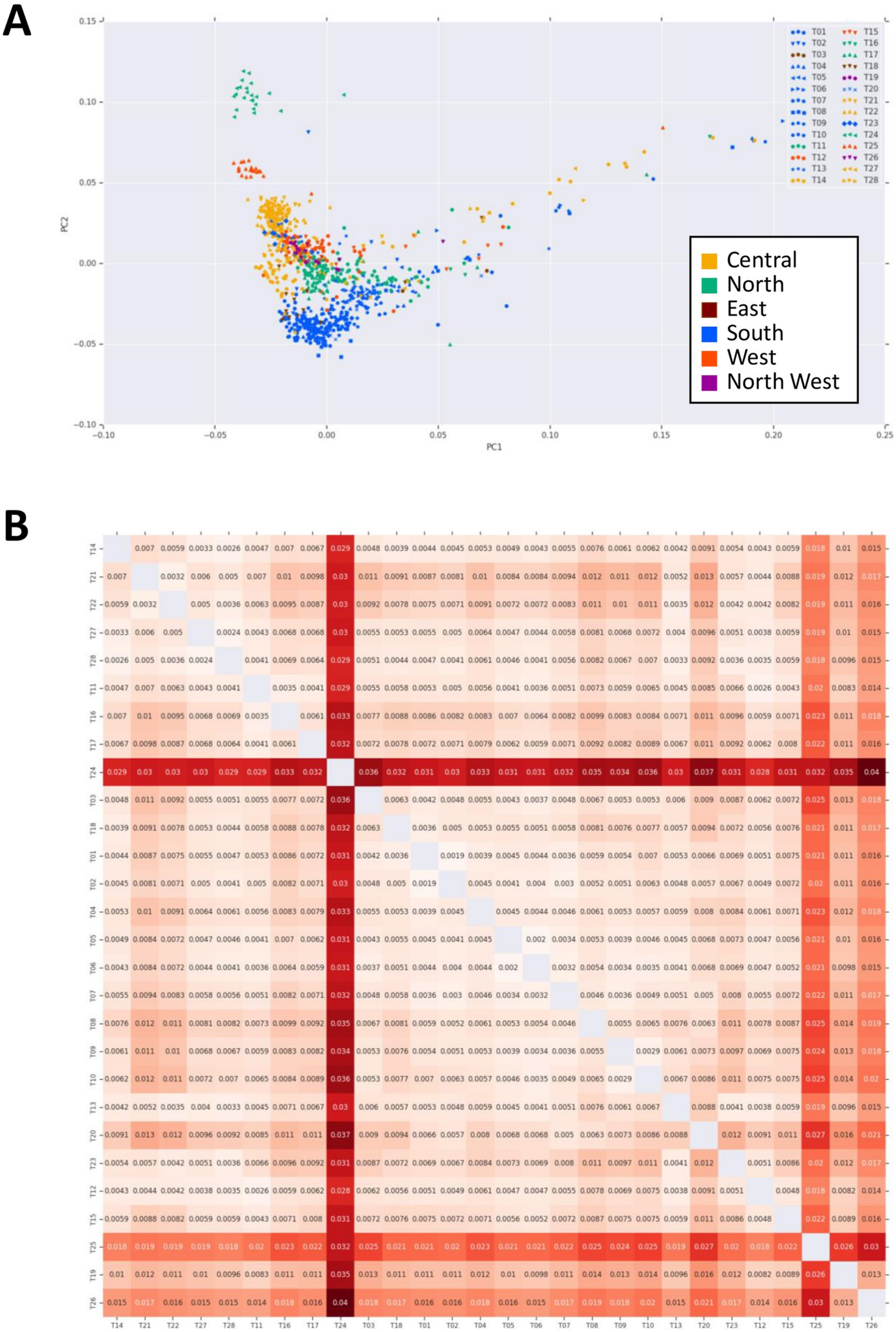
ranging from 12.7% for Europe to 0.60% for Oceania (Fig 1A). Interestingly, we observed a long tail with respect to both PC1 and PC2 towards the African population cluster (S1 Fig). To overcome the possible bias caused by the differences of sample sizes, we repeated the same analysis using similar and smaller sample sizes for those regions and the results of subsampling were compatible (S14 Fig), suggesting this was not a major factor. Instead, the wide variance and well-known historical interactions between Saudi Arabia and Africa in the slave trade argue for a recent admixture between these two populations [20]. This notion is corroborated by the inverse correlation between degree of admixture and estimated date of admixture (see below).

Arabia is well known to be a focal point of human interaction throughout human history. In order to explore the influence of admixture spatially across the peninsula we grouped the tribes by geographic location and compared estimated admixture proportions for intercontinental admixture and the clustering when K = 7 (Fig 1A). While some genetic influence from major continental populations is evident in each group, large population groups near the center were observed to have substantially less admixture compared to population groups in the Southern, Eastern and Northern parts of the country (Mann-Whitney U test with Bonferroni correction, $p < 0.05$; S1 Table). Intercontinental admixture primarily derives from Europe and Africa, while sources of American and Oceanian admixture are minor, and sporadic (Fig 1A). Interestingly, the minimum squared error admixture (K = 18) shows several ancestral components specific to the Saudi Arabian samples (Fig 1B and 1C).

Signature of tribalism in Arabia

Population divergence showed a clear pattern of clustering along tribal lines, with some overlap that can be explained through geographic proximity and historical records of intermarriage (Figs 1, 2, S15 and S16). Hierarchical F_{ST} for regional, tribal and individual levels are examined (S2 Table), suggesting the regional difference is relatively low. For example, populations T01 and T02 both derive from the south-west region of Saudi Arabia, where no physical boundaries prevent interaction with neighboring tribes. Similarly, populations T21 and T22 co-inhabit the Central region, and T11 and T17 the Northern region of Saudi Arabia and are known to intermarry, which explains the gene flow within these pairs. On the other hand, the most distinct from the other tribes in Admixture analysis, T24, is distinct even from other tribes in the northern region of Arabia where it resides, consistent with historical accounts that this tribe has seen very little marriage with other tribes for many generations. We then examined haplogroups of Y chromosome and mitochondrial genome in these indigenous Arab tribes (S17 and S18 Figs). For haplogroups in Y chromosome, though the major component was J1 haplogroup which is known to be widely distributed among Middle Eastern populations [21], there are tribal-specific patterns such as a higher ratio of E1 haplogroup in some tribes. This haplogroup is reported dominantly in African region [21], suggesting the admixture event shown in Y Chromosome. For mitochondria, the composition of haplogroups among tribes was more diverged and global than that in Y chromosome. We could observe that H2 haplogroups were dominant, which are found in Eastern Europe, Middle and Near East [22]. Also, some tribes showed African L haplogroups. Tribal compositions of haplogroups in Y chromosome and mitochondrial genome were not consistent, suggesting the distinct admixture scenarios between males and females. As the composition of haplogroups in Y chromosome is conservative, males in each tribe tend not to move across tribes or beyond tribes.

While we observe outliers in multiple tribes, the samples tend to coincide with populations from the denser sections of the principal component space (Figs 2A and S19).



	T14	T21	T22	T27	T28	T11	T16	T17	T24	T03	T18	T01	T02	T04	T05	T06	T07	T08	T09	T10	T13	T20	T23	T12	T15	T25	T19	T26
T14		0.007	0.0059	0.0033	0.0026	0.0047	0.007	0.0067	0.029	0.0048	0.0039	0.0044	0.0045	0.0053	0.0049	0.0043	0.0055	0.0076	0.0061	0.0062	0.0042	0.0091	0.0054	0.0043	0.0059	0.018	0.01	0.015
T21	0.007		0.0032	0.006	0.005	0.007	0.01	0.0098	0.03	0.011	0.0091	0.0087	0.0081	0.01	0.0084	0.0084	0.0094	0.012	0.011	0.012	0.0052	0.013	0.0057	0.0044	0.0088	0.019	0.012	0.017
T22	0.0059	0.0032		0.005	0.0036	0.0063	0.0095	0.0087	0.03	0.0092	0.0078	0.0075	0.0071	0.0091	0.0072	0.0072	0.0083	0.011	0.01	0.011	0.0035	0.012	0.0042	0.0042	0.0082	0.019	0.011	0.016
T27	0.0033	0.006	0.005		0.0024	0.0043	0.0068	0.0068	0.03	0.0055	0.0053	0.0055	0.005	0.0064	0.0047	0.0044	0.0058	0.0081	0.0068	0.0072	0.004	0.0096	0.0051	0.0038	0.0059	0.019	0.01	0.015
T28	0.0026	0.005	0.0036	0.0024		0.0041	0.0069	0.0064	0.029	0.0051	0.0044	0.0047	0.0041	0.0061	0.0046	0.0041	0.0056	0.0082	0.0067	0.007	0.0033	0.0092	0.0036	0.0035	0.0059	0.018	0.0096	0.015
T11	0.0047	0.007	0.0063	0.0043	0.0041		0.0035	0.0041	0.029	0.0055	0.0058	0.0053	0.005	0.0056	0.0041	0.0036	0.0051	0.0073	0.0059	0.0065	0.0045	0.0085	0.0066	0.0026	0.0043	0.02	0.0083	0.014
T16	0.007	0.01	0.0095	0.0068	0.0069	0.0035		0.0061	0.033	0.0077	0.0088	0.0086	0.0082	0.0083	0.007	0.0064	0.0082	0.0099	0.0083	0.0084	0.0071	0.011	0.0096	0.0059	0.0071	0.023	0.011	0.018
T17	0.0067	0.0098	0.0087	0.0068	0.0064	0.0041	0.0061		0.032	0.0072	0.0078	0.0072	0.0071	0.0079	0.0062	0.0059	0.0071	0.0092	0.0082	0.0089	0.0067	0.011	0.0092	0.0062	0.008	0.022	0.011	0.016
T24	0.029	0.03	0.03	0.03	0.029	0.029	0.033	0.032		0.036	0.032	0.031	0.03	0.033	0.031	0.031	0.032	0.035	0.034	0.036	0.03	0.037	0.031	0.028	0.031	0.032	0.035	0.04
T03	0.0048	0.011	0.0092	0.0055	0.0051	0.0055	0.0077	0.0072	0.036		0.0063	0.0042	0.0048	0.0055	0.0043	0.0037	0.0048	0.0067	0.0053	0.0053	0.006	0.009	0.0087	0.0062	0.0072	0.025	0.013	0.018
T18	0.0039	0.0091	0.0078	0.0053	0.0044	0.0058	0.0088	0.0078	0.032	0.0063		0.0036	0.005	0.0053	0.0055	0.0051	0.0058	0.0081	0.0076	0.0077	0.0057	0.0094	0.0072	0.0056	0.0076	0.021	0.011	0.017
T01	0.0044	0.0087	0.0075	0.0055	0.0047	0.0053	0.0086	0.0072	0.031	0.0042	0.0036		0.0019	0.0039	0.0045	0.0044	0.0036	0.0059	0.0054	0.007	0.0053	0.0066	0.0069	0.0051	0.0075	0.021	0.011	0.016
T02	0.0045	0.0081	0.0071	0.005	0.0041	0.005	0.0082	0.0071	0.03	0.0048	0.005	0.0019		0.0045	0.0041	0.004	0.003	0.0052	0.0051	0.0063	0.0048	0.0057	0.0067	0.0049	0.0072	0.02	0.011	0.016
T04	0.0053	0.01	0.0091	0.0064	0.0061	0.0056	0.0083	0.0079	0.033	0.0055	0.0053	0.0039	0.0045		0.0045	0.0044	0.0046	0.0061	0.0053	0.0057	0.0059	0.008	0.0084	0.0061	0.0071	0.023	0.012	0.018
T05	0.0049	0.0084	0.0072	0.0047	0.0046	0.0041	0.007	0.0062	0.031	0.0043	0.0055	0.0045	0.0041	0.0045		0.002	0.0034	0.0053	0.0039	0.0046	0.0045	0.0068	0.0073	0.0047	0.0056	0.021	0.01	0.016
T06	0.0043	0.0084	0.0072	0.0044	0.0041	0.0036	0.0064	0.0059	0.031	0.0037	0.0051	0.0044	0.004	0.0044	0.002		0.0032	0.0054	0.0034	0.0035	0.0041	0.0068	0.0069	0.0047	0.0052	0.021	0.0098	0.015
T07	0.0055	0.0094	0.0083	0.0058	0.0056	0.0051	0.0082	0.0071	0.032	0.0048	0.0058	0.0036	0.003	0.0046	0.0034	0.0032		0.0046	0.0036	0.0049	0.0051	0.005	0.008	0.0055	0.0072	0.022	0.011	0.017
T08	0.0076	0.012	0.011	0.0081	0.0082	0.0073	0.0099	0.0092	0.035	0.0067	0.0081	0.0059	0.0052	0.0061	0.0053	0.0054	0.0046		0.0055	0.0065	0.0076	0.0063	0.011	0.0078	0.0087	0.025	0.014	0.019
T09	0.0061	0.011	0.01	0.0068	0.0067	0.0059	0.0083	0.0082	0.034	0.0053	0.0076	0.0054	0.0051	0.0053	0.0039	0.0034	0.0036	0.0055		0.0029	0.0061	0.0073	0.0097	0.0069	0.0075	0.024	0.013	0.018
T10	0.0062	0.012	0.011	0.0072	0.007	0.0065	0.0084	0.0089	0.036	0.0053	0.0077	0.007	0.0063	0.0057	0.0046	0.0035	0.0049	0.0065	0.0029		0.0067	0.0086	0.011	0.0075	0.0075	0.025	0.014	0.02
T13	0.0042	0.0052	0.0035	0.004	0.0033	0.0045	0.0071	0.0067	0.03	0.006	0.0057	0.0053	0.0048	0.0059	0.0045	0.0041	0.0051	0.0076	0.0061	0.0067		0.0088	0.0041	0.0038	0.0059	0.019	0.0096	0.015
T20	0.0091	0.013	0.012	0.0096	0.0092	0.0085	0.011	0.011	0.037	0.009	0.0094	0.0066	0.0057	0.008	0.0068	0.0068	0.005	0.0063	0.0073	0.0086	0.0088		0.012	0.0091	0.011	0.027	0.016	0.021
T23	0.0054	0.0057	0.0042	0.0051	0.0036	0.0066	0.0096	0.0092	0.031	0.0087	0.0072	0.0069	0.0067	0.0084	0.0073	0.0069	0.008	0.011	0.0097	0.011	0.0041	0.012		0.0051	0.0086	0.02	0.012	0.017
T12	0.0043	0.0044	0.0042	0.0038	0.0035	0.0026	0.0059	0.0062	0.028	0.0062	0.0056	0.0051	0.0049	0.0061	0.0047	0.0047	0.0055	0.0078	0.0069	0.0075	0.0038	0.0091	0.0051		0.0048	0.018	0.0082	0.014
T15	0.0059	0.0088	0.0082	0.0059	0.0059	0.0043	0.0071	0.008	0.031	0.0072	0.0076	0.0075	0.0072	0.0071	0.0056	0.0052	0.0072	0.0087	0.0075	0.0075	0.0059	0.011	0.0086	0.0048		0.022	0.0089	0.016
T25	0.018	0.019	0.019	0.019	0.018	0.02	0.023	0.022	0.032	0.025	0.021	0.021	0.02	0.023	0.021	0.021	0.022	0.025	0.024	0.025	0.019	0.027	0.02	0.018	0.022		0.026	0.03
T19	0.01	0.012	0.011	0.01	0.0096	0.0083	0.011	0.011	0.035	0.013	0.011	0.011	0.011	0.012	0.01	0.0098	0.011	0.014	0.013	0.014	0.0096	0.016	0.012	0.0082	0.0089	0.026		0.013
T26	0.015	0.017	0.016	0.015	0.015	0.014	0.018	0.016	0.04	0.018	0.017	0.016	0.016	0.018	0.016	0.015	0.017	0.019	0.018	0.02	0.015	0.021	0.017	0.014	0.016	0.03	0.013	

Fig 2. Genetic differentiation of indigenous Arab population. (A) Principal component analysis of 957 indigenous Arab samples. Tribal affiliations for T01-T28 are represented by different symbols. Colors correspond to the geographical location of Arabian Peninsula. (B) Pairwise F_{ST} values among indigenous Arab tribes. Values at each combination show the pairwise F_{ST}, calculated by BEDASSLE package in R. Colored intensity indicates the degree of pairwise F_{ST} at each combination. Brighter red indicates more differentiated between tribes compared.

https://doi.org/10.1371/journal.pgen.1009210.g002

To assess the population divergence, we estimated pairwise F_{ST} values between all tribes (Fig 2B). The greatest distance measured between any two groups was 0.04 (T24 and T26). Much of this divergence was due to the existence of two outlier populations (T24 and T25). These two outlier tribes have the highest average of inbreeding coefficients among other tribes, consistent with historical tradition of strict intratribal marriage within these two tribes (Fig 3 and S3 Table). Among the rest of the tribes, we observe a wide gradient of divergence. Interestingly, we observe some near 0 pairwise F_{ST} estimates between some tribes e.g. 0.0024 (T27 and T28), 0.0032 (T21 and T22) and 0.0041 (T11 and T17), arguing for a close interaction between them, again consistent with historical tradition of intermarriage between these specific tribes on the basis of perceived "compatibility" of lineages. To further examine outcomes of inbreeding, Runs of Heterozygosity (ROH) analysis was performed (S20 Fig). It showed that T24 and T25 have the highest median of sum length of ROH (S20A Fig), consistent with above mentioned. Possible scenarios of the higher inbreeding coefficient are recent inbreeding or past bottleneck effects. We plotted tribes at the sum total length of ROH (SROH) versus the total number of ROH (NROH) of individuals (S20B Fig). Recent inbreeding tends to increase SROH as longer ROHs are maintained, whereas bottleneck effect tends to increase NROH with depletion of long ROH [23]. As shown in S20B Fig, the distribution of tribes is diverse with a tendency to cluster at the center, suggesting recent inbreeding events for those tribes. A notable exception is observed for T24 and T25, however, which display an extreme shift caused by the large values of both SROH and NROH. Therefore, in addition to the recent practice of consanguinity observed in many other tribes, these two tribes show a strong evidence of a past population bottleneck. We also traced the demographic history by estimating effective population size (N_e) of each tribe (S21 Fig). Most of tribes showed increasing N_e gradually over generations. T24 and T25 seem to have a relative decrease of N_e around 25 generations ago, implying the past bottleneck event for these tribes. It is noteworthy that the increasing trend of T10 is a bit drastic, compared to the other tribes, indicating the contribution of extensive gene flow or admixture events in T10.

When we ran admixture analysis on only the Saudi Arabian populations, we find a minimum squared error associated with five ancestral populations K = 5 (S15 Fig). When comparing these admixture components with tribes, we observed a clear geographic pattern to the various sources and further evidence of rampant migration and interaction between ancestral populations. The previously observed outlier, T24, appears to be a near homogeneous representation of one of these populations, which has admixed with nearly every other tribe to varying degrees. Similarly, we can observe a potential gradient of gene flow from the southern point of the peninsula.

Patterns of human migration and drift were recapitulated using TreeMix on the 28 tribes with reference populations (Fig 4). The derived tree shows limited drift among tribes. ALDER was used to estimate the date of African admixture for each tribe as shown in S4 Table and plotted in Fig 5A. We observed that the estimated number of generations ranged 11 to 41 with an average mixture date of around 25 generations ago. This seemed to be relatively recent as human history in Arabian Peninsula can be predated to $>$ 85,000 years ago [10]. This may be due to the limitation of our platform or the traits are difficult to trace back as they are too ancient. We also used *f4-ratio estimate* to quantify the African proportions in the 28 tribes and

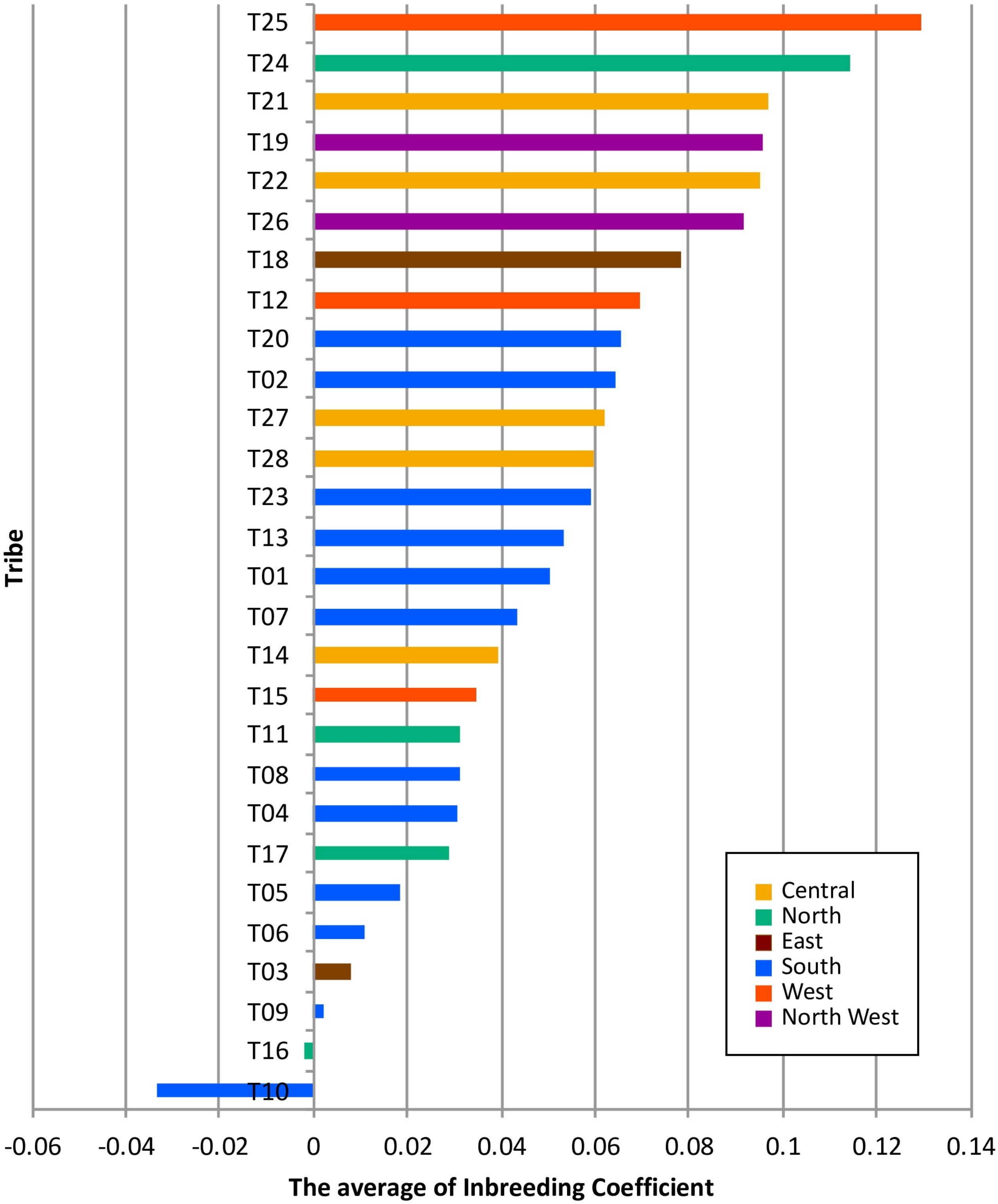


Fig 3. The average of inbreeding coefficients for 28 tribes. The bar chart shows the average of inbreeding coefficient for each tribe shown in decreasing order. The color indicates the regional group in the Arabian Peninsula. The inbreeding coefficient for each individual was calculated as "Fhat2" estimate using PLINK.

https://doi.org/10.1371/journal.pgen.1009210.g003

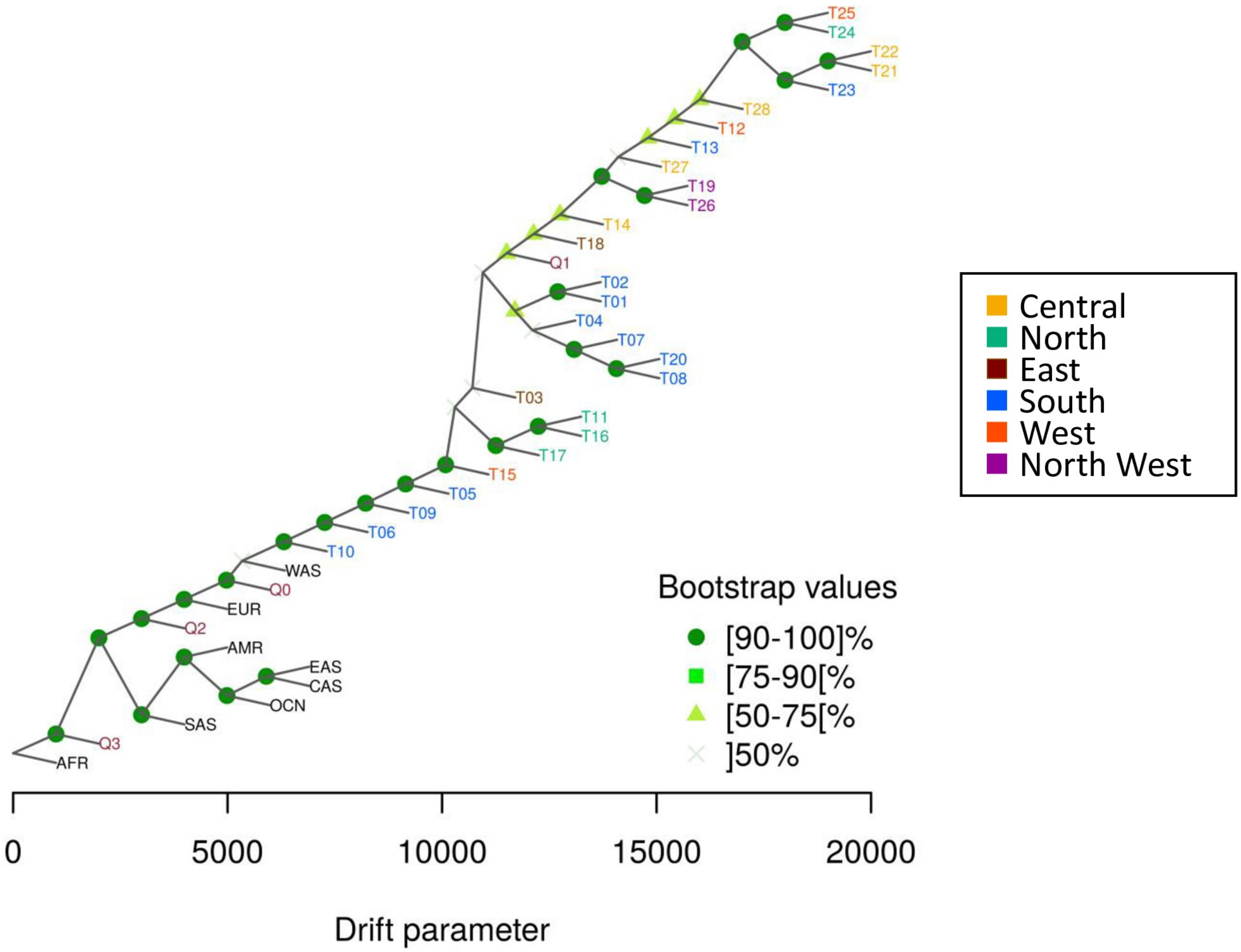


Fig 4. TreeMix representation of indigenous Arab populations with reference populations. Saudi tribes (T01-T29) are coded by regional colors. Bootstrap confidence value is indicated by distinct symbols at each node. AFR; Africa, AMR; America, CAS; Central Asia, EAS; East Asia, EUR; Europe, OCN; Oceania, SAS; South Asia, WAS; West Asia, Q1; Bedouin in Qatar genome project, Q2; Persian-South Asian in Qatar genome project, Q3; African in Qatar genome project, Q0; unassigned in Qatar genome project.

https://doi.org/10.1371/journal.pgen.1009210.g004

the result are plotted in Fig 5B. Most tribes have inherited 7–16% African ancestry, an estimate compatible with that reported for Near Eastern populations (Bedouin and Palestinian) in [24]. Demographic history of T10 was observed uniquely as above (S21 Fig), and this T10 showed the highest ratio of African ancestry. It suggests that there is the anthropogenic relationship between Saudi tribes and African population. Considering the outcome of both analyses, ALDER and f4-ratio estimate, we found that tribes with higher African ancestry tend to have a recent admixture event. In other words, those with the highest African admixture are those with the shortest generation time from the date of admixture, consistent with the previously stated notion that this African admixture is likely a relatively recent event caused by the slave trade [20].

Discussion

Patients of Arab descent with genetic disorders have long been instrumental in the identification of autosomal recessive disease genes by virtue of positional mapping and, more recently, next generation sequencing [12,25,26]. Despite this major contribution to the global efforts of

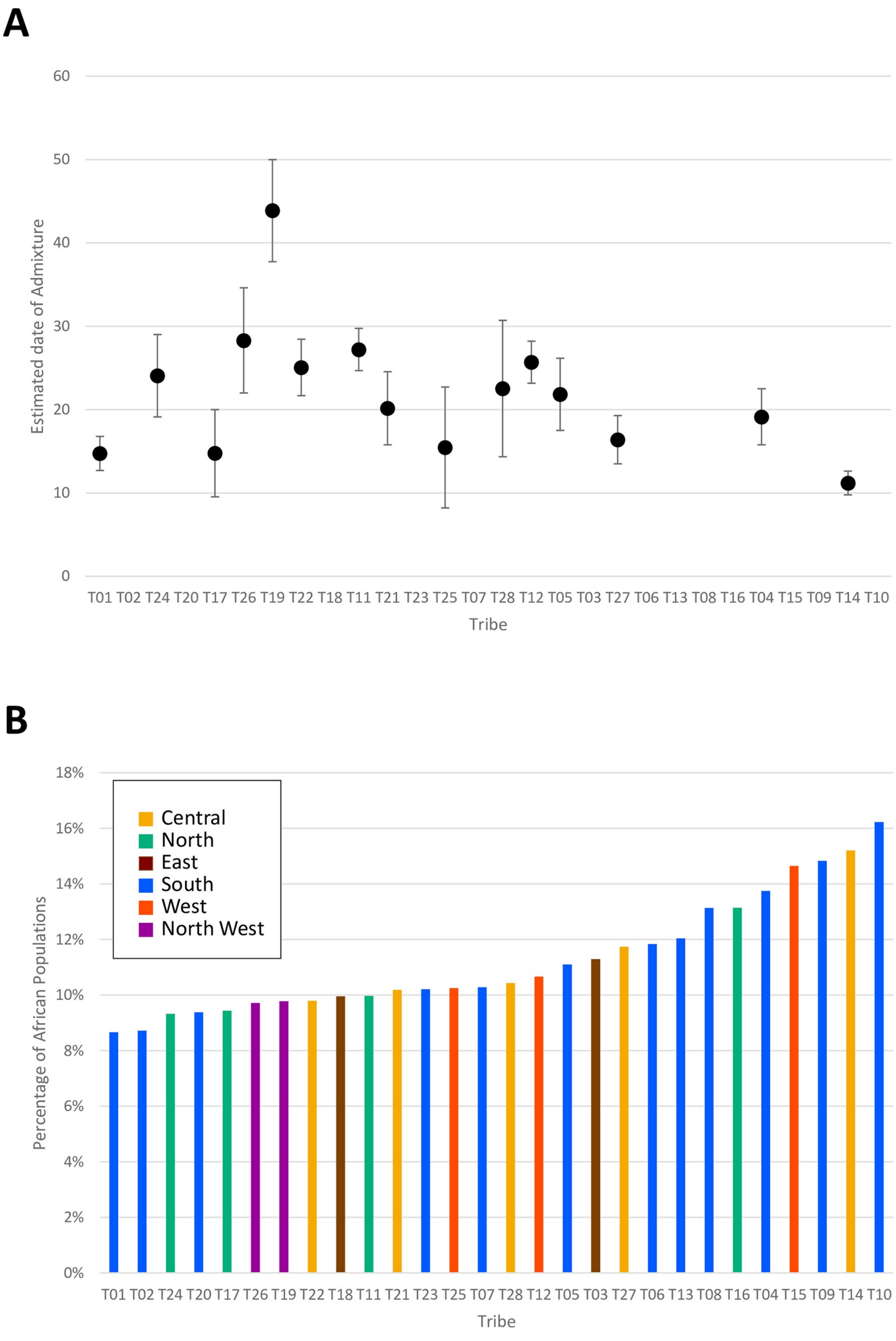

A
60
50
40
30
20
10
0
Estimated date of Admixture
T01 T02 T24 T20 T17 T26 T19 T22 T18 T11 T21 T23 T25 T07 T28 T12 T05 T03 T27 T06 T13 T08 T16 T04 T15 T09 T14 T10
Tribe
B
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
Percentage of African Populations
Central
North
East
South
West
North West
T01 T02 T24 T20 T17 T26 T19 T22 T18 T11 T21 T23 T25 T07 T28 T12 T05 T03 T27 T06 T13 T08 T16 T04 T15 T09 T14 T10
Tribe

Fig 5. (A) The estimated date of African admixture, (B) The ratio of African ancestry for 28 tribes. Tribes are sorted from right to left by decreasing proportion of estimated African admixture.

https://doi.org/10.1371/journal.pgen.1009210.g005

medical annotation of the human genome, little is known about their variome, much less population stratification. We have recently studied a large number of exomes from the "Greater Middle East" and shown the tremendous value of this information in informing novel disease gene discovery [13]. In that study, we have also shown that the degree of variation in genomes from the Middle East is intermediary between Africans and the other world populations, consistent with the notion that the Middle East was inhabited for an extended period of time shortly after the out of Africa migration. However, we did not focus in that study on Arabian Peninsula although we did show the important contribution of the Arabian Peninsula to the genomes of nearly all Greater Middle Eastern populations, in line with the history of Arab exodus in the 7th century at the time of Islamic expansion [13].

In this study, we show that the generic reference to natives of Arabia as "Bedouins" overshadows important population stratification along tribal lines. Although modernization and relaxed social norms will likely change this in the future, previous work suggests that founder mutations are typically only shared between tribes as a function of being very old because they predate splitting of tribes, rather than due to recent mixing [27]. Therefore, the effect of tribalism on the genetic landscape in Arabia should not only be considered for its historical legacy but also for its influence on modern day descendants. In particular, the need for a local variome database that is inclusive of all tribes cannot be overemphasized. Indeed, we have encountered numerous reports of likely pathogenic variants from clinical sequencing labs that were subsequently shown to represent local variants that may only be common in individual tribes (the value of the local variome to enhance specificity of mutation calling for Mendelian diseases will be discussed elsewhere) [28].

The local population structure has been conducive to genetic studies that have Mendelian focus. This in part explains the dearth of GWAS from this region. It is inevitable, however, that GWAS will be pursued more aggressively in the future to address the genetic risk of complex diseases, and a good understanding of the local structure is a prerequisite for this. We also speculate that the approach in Sardinia, which is an isolated population that accumulated rare variants [29], can be applicable to indigenous Arab populations and thus it might be useful to future development of GWAS design for this population. Rodriguez-Flores and colleagues have recently studied the genomes of 56 indigenous individuals from Arabia and found a higher degree of mixing with the Neanderthals than Africans but lower than Europeans and Asians [19]. These results, when combined with our previous data that indigenous Arabs have lower degree of genetic diversity than Africans but higher than Europeans and Asians, strongly support the notion that Arabs are the descendants of the first wave of out of Africa humans who later migrated further to Asia and Europe [13]. In this study, we also confirm the unique clustering of indigenous Arabs compared to the Asians and Europeans as shown by Rodriguez-Flores *et al* [19].

The much larger sample size and the careful selection of tribal affiliation allowed us in this study to analyze the population structure of indigenous Arabs at an unprecedented resolution. It is remarkable that self-identified tribal origin was highly predictive of the individual's place in the cluster. Further support of the accuracy of our data comes from the finding that tribes that shared the same geographic location had a shorter genetic distance compared to those that historically concentrated in a different locale. Apart from the forensic and anthropological applications of these findings, we posit that they also impact healthcare. For example, the finding that most founder recessive mutations are tribe-specific points to the limited benefit of educational efforts that focus on consanguinity and fail to discuss intra-tribal unions [30]. From a sociological point of view, the finding of several ancient founder variants that are shared across tribes should provide tangible evidence of the shared heritage of all these tribes such that the newer generations are better informed when deciding on out-of-tribe unions.

In conclusion, we report the first genetic analysis of the tribal structure of Arabia and show that this ancient legacy has significant relevance to contemporary genomic medicine. We hope that the data from our analysis will contribute to filling the current gap in our understanding of the variome of this part of the world.

Materials and methods

Ethics statement

The source of human subjects in this study is a large collection of IRB-approved research protocols that involve the genetics of various genetic diseases under approval numbers: 2121053, 2070023, 208006 for KFSH&RC. This study was also approved by KAUST under 15IBEC39. Written informed consent was obtained from all subjects. We selected 1,073 subjects for whom genotyping as well as self-stated tribal affiliations are available. There were 28 tribes in total representing >95% of self-identified tribes in our database. Tribes are anonymized except for its geographical location (S5 Table).

Genotyping and data preprocessing

532,615 autosomal SNPs were genotyped for a sample of 1,073 individuals using the Affymetrix Axiom genotyping assay (Axiom Genome-wide CEU 1 Array Plate, AxiomGWH-96Array, Axiom 2.0 Kit). Sample preparation including whole genome amplification, fragmentation, denaturation and hybridization were all performed according to manufacturer's specifications and recommendations (Affymetrix, Santa Clara, California, USA). Automated, high-throughput processing of genome-wide SNP genotyping was carried out using the GeneTitan system (Affymetrix).

The relatedness was assessed using kinship coefficients estimated by KING [31]. We ran KING to extract a list of individuals that contains no pairs of individuals with a first-, second-, or third-degree relationship. PLINK [32,33] was used to prune the 532,615 autosomal SNPs down to 455,266 SNPs with a minor allele frequency greater than 1%, a missing rate less than 10% and a Hardy-Weinberg equilibrium (HWE) deviation p-value of no less than 0.01. We identified individuals who have an extreme low Z score (less than 4 standard deviation units) as outliers by PLINK outlier detection diagnostics and excluded them from subsequent analysis. Only the remaining 957 unrelated individuals were used in the subsequent analysis (S3 Table), including PCA, Wright's fixation index (F_{ST}) measurement, admixture analysis, TreeMix analysis, inbreeding coefficient, and estimating date and degree of admixture using ALDER and f4-ratio estimation.

The 1000 Genomes Project [15], Human Genome Diversity Project (HGDP) [34], The Simons Genome Diversity Project (SGDP) [35], and Qatari Genome [19] data were used as a reference to assess how the Saudi population samples related to other human populations. Same as Saudi data, we used KING program to exclude duplicated individuals form integrated reference data. Because the Saudi samples and samples in reference databases were analyzed on different platforms, analysis was limited to the intersection of SNPs between these platforms. The intersection contained 426,056 SNPs, which were sufficient to produce reliable results and were used for subsequent analysis.

Determination of population structure and admixture

PCA was performed using PLINK. We ran PCA on the Saudi samples and plotted all the samples onto the first two principal components. F_{ST} was calculated using the R BEDASSLE package [36] to explore the degree of differentiation between tribes. Hierarchical F_{ST} tests (AMOVA for SNP dataset) was performed by the R hierfstat package [37].

To compare the Saudi samples with other continental populations, PCA was performed using PLINK and we plotted the result by our custom python script. We performed Identity-by-Descent (IBD)-based hierarchical clustering.

Admixture analysis was conducted using ADMIXTURE [38] on the combined dataset of 1,672 samples representing 957 Saudi samples and 3,691 reference samples that represent the nine regions Africa, Europe, Central Asia, East Asia, South Asia, West Asia, Oceania, America, and Qatar (S6 Table). ADMIXTURE was run using default settings with the cross-validation procedure. Minimum squared error values calculated from the cross-validation procedure in ADMIXTURE to evaluate a good value of the number of ancestral populations K.

To evaluate the influence of differences of population sizes, we performed ADMIXTURE with downsized data set. We subsampled 50 individuals randomly (or all individuals if the number of individuals in the population is less than 50) from each population, then performed ADMIXTURE for the downsized data set.

TreeMix analysis

We performed a TreeMix analysis [39] of the 28 tribes with reference populations (S6 Table) with default settings. Bootstrap with 1,000 replications for TreeMix was performed by BITE R package [40] and the consense program of PHYLIP version 3.6 [41].

Inbreeding coefficient

The inbreeding coefficient was calculated for the Saudi samples using "*fhat2*" estimate of PLINK 1.9 [33] in order to compare the rate of endogamy in the different tribes as a proxy of their degree of isolation (tribes with the least degree of intermixing will have the highest inbreeding coefficient). The average and standard deviation for each tribe was calculated. The average values of inbreeding coefficient of 28 tribes are shown in a bar plot where tribes are sorted from top to bottom in a decreasing order.

Estimating date and degree of admixture

The date of admixture for each tribe was estimated using ALDER 1.2 [42]. As we need detailed population data for these analyses, we focused on only HGDP dataset among integrated reference datasets. The populations "Nigeria (Yoruba)" and France from HGDP dataset was used as two reference populations. F4-ratio estimation was used to estimate the proportion of African ancestry in the 28 tribes as performed in [24]. To estimate the proportion of African ancestry in $tribe_X$, *f4*-ratio estimate is obtained by computing the ratio of two (*f4*) statistics, as follow:

$$f_4 - ratio\ estimate\ (tribe_X) = \frac{f4\ (San, China; tribe_X, France)}{f4\ (San, China; Yoruba, France)}$$

where $tribe_X$ is one of the 28 tribes and San, China, France, Nigeria (Yoruba) from HGDP dataset were used as reference populations. The *f4*-ratio estimate is proportional to the amount of African-mixture proportion in $tribe_X$.

Runs of homozygosity

The runs of homozygosity (ROH) was calculated for the Saudi samples by PLINK 1.9 [29]. To show the distribution of ROH, we created a scatter plot between means of the sum total length of ROH (SROH) and the total number of ROH (NROH) of individuals. The violin plot of SROH for tribes is also created.

Effective population size

We estimated N_e based on long segments of IBD by the IBDNe program [43]. To run the IBDNe program, we prepared imputed and phased data by the Beagle 5.1 [44,45], and detected IBD segments by the hap-ibd program [46]. We plotted the Ne from four generations to 50 generations ago [43].

Assignment Y chromosome and mitochondrial Haplogroup

For haplogroup assignments, we employed the Yfitter v0.3 [47] for Y chromosome haplogroup and the HaploGrep v2.2.0 [48] for mitochondrial haplogroup. Both programs assign known haplotype to Saudi samples.

Supporting information

S1 Fig. Principal component analysis of indigenous Arab with reference populations.
(TIF)

S2 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting Saudi tribes, Qatari populations (QTR; Q0-Q3) and West Asian population (WAS; including Bedouin).
(TIF)

S3 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting African populations (AFR).
(TIF)

S4 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting American populations (AMR).
(TIF)

S5 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting Central Asian populations (CAS).
(TIF)

S6 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting East Asian populations (EAS).
(TIF)

S7 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting European populations (EUR).
(TIF)

S8 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting Saudi tribes.
(TIF)

S9 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting Oceanian populations (OCN).
(TIF)

S10 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting Qatari populations (QTR).
(TIF)

S11 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting South Asia populations (SAS).
(TIF)

S12 Fig. Principal component analysis of indigenous Arab with reference populations, highlighting West Asia populations (WAS).
(TIF)

S13 Fig. ADMIXTURE analysis of indigenous Arab and reference populations. Results of ADMIXTURE analysis for 4,648 samples representing 957 Saudi samples and 3,691 reference samples. 25 iterations of K were run, from 1 to 25, to optimize clustering. The results between K = 16 and 20 are shown (K = 18 is the optimal one: see main text). Each vertical bar represents a single individual. The y axis shows the estimated proportion of the genome assigned to each ancestral cluster.
(TIF)

S14 Fig. ADMIXTURE analysis of indigenous Arab and reference populations by subsampling. 50 individuals are randomly subsampled from each representative population. Abbreviations of populations are the same as in S6 Table. (A) Cross-validation error for K runs from 2 to 15. K = 7 has the lowest cross-validation error. (B) Results of ADMIXTURE analysis at K = 7. The y axis shows the estimated proportion of the genome assigned to each ancestral cluster.
(TIF)

S15 Fig. ADMIXTURE analysis of indigenous Arab populations. (A) Results of ADMIXTURE analysis for 957 Saudi samples across the 28 Tribes. Nine iterations of K were run, from 1 to 9, to optimize clustering. The results between K = 3 and 5 are shown. Each vertical bar represents a single individual. The y axis shows the estimated proportion of the genome assigned to each ancestral cluster. (B) Cross-validation error for K runs from 1 to 9. K = 5 has the lowest cross-validation error.
(TIF)

S16 Fig. IBD-based hierarchical clustering. (A) Indigenous Arab populations and reference populations. (B) Indigenous Arab populations.
(TIF)

S17 Fig. Haplogroup assignment for (A) Y chromosome and (B) mitochondrial genome.
(TIF)

S18 Fig. Haplogroup assignment of indigenous Arab tribes and their geographical location. (A) Y chromosome and (B) mitochondrial genome. The map was made with Natural Earth (public domain).
(TIF)

S19 Fig. Principal component analysis of indigenous Arab samples. This plot is corresponding to Fig 2A. Tribal affiliations for T01-T28 are represented by different symbols. Colors correspond to the geographical location of Arabian Peninsula as in Fig 2A. Average of PC1 and

PC2 for each tribe is plotted by the tribal symbol and color, and standard deviation is represented by lines. Color of each sample is changed to gray.
(TIF)

S20 Fig. Trends of Runs of Homozygosity for indigenous Arab tribes. (A) Violin plot of sum length of runs of homozygosity (SROH) in Indigenous Arab tribes. Color correspond to the geographical group, and tribes are sorted by this group. The violin shows a colored kernel density trace with the interquartile range as a black line and the median as a white circle. (B) Scatter plot of the mean SROH and number of ROH (NROH). Symbols and colors are corresponding to tribes.
(TIF)

S21 Fig. Estimation of effective population size for Indigenous Arab tribes. Effective population size (N_e) is estimated by IBDNe. Colors and symbols are corresponding to each tribe. Series of estimated N_e from 4 to 50 generations is plotted.
(TIF)

S1 Table. Statistical test for regional differences of non-Arab ancestries.
(PDF)

S2 Table. Hierarchical F_{ST} for Indigenous Arab populations.
(PDF)

S3 Table. Inbreeding coefficients of 28 Saudi tribes.
(PDF)

S4 Table. The estimated African proportions and date of admixture for 28 Saudi tribes.
(PDF)

S5 Table. Summary of the Saudi dataset.
(PDF)

S6 Table. Summary of the reference dataset.
(PDF)

S1 Data. Genotype data for each tribe.
(GZ)

Acknowledgments

We thank the study participants for their enthusiastic participation. We also thank the Genotyping and Sequencing Core Facilities at KFSH&RC for their technical support. We are grateful to Mais Hashem, Niema Ibrahim and Firdous Abdulwahab for their help as research coordinators and to Reem Alsuwailem for her help in data collection. We are extremely grateful to Prof. Rodriguez-Flores and his colleagues at Cornell University for their generosity with sharing the Qatari Genome data.

Author Contributions

Investigation: Katsuhiko Mineta, Kosuke Goto.

Supervision: Takashi Gojobori, Fowzan S. Alkuraya.

Writing – review & editing: Katsuhiko Mineta, Kosuke Goto, Takashi Gojobori, Fowzan S. Alkuraya.

References

1. Przeworski M, Hudson RR, Di Rienzo A. Adjusting the focus on human variation. Trends in Genetics. 2000; 16(7):296–302. https://doi.org/10.1016/s0168-9525(00)02030-8 PMID: 10858659
2. Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, et al. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature genetics. 2002; 32 (1):135–42. https://doi.org/10.1038/ng947 PMID: 12161752
3. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics. 2009; 10(4):241–51. https://doi.org/10.1038/nrg2554 PMID: 19293820
4. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Science translational medicine. 2011; 3 (65):65ra4–ra4. https://doi.org/10.1126/scitranslmed.3001756 PMID: 21228398
5. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic misdiagnoses and the potential for health disparities. New England Journal of Medicine. 2016; 375(7):655–65. https://doi.org/10.1056/NEJMsa1507092 PMID: 27532831
6. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019; 570:514–8. https://doi.org/10.1038/s41586-019-1310-4 PMID: 31217584
7. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nature communications. 2019; 10(1):1–9. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773
8. Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nature communications. 2012; 3:698. https://doi.org/10.1038/ncomms1701 PMID: 22426219
9. Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, et al. The ancestry and affiliations of Kennewick Man. Nature. 2015; 523:455–8. https://doi.org/10.1038/nature14625 PMID: 26087396
10. Groucutt HS, Grün R, Zalmout IA, Drake NA, Armitage SJ, Candy I, et al. Homo sapiens in Arabia by 85,000 years ago. Nature ecology & evolution. 2018; 2(5):800–9. https://doi.org/10.1038/s41559-018-0518-2 PMID: 29632352
11. Fernandes V, Alshamali F, Alves M, Costa MD, Pereira JB, Silva NM, et al. The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. The American Journal of Human Genetics. 2012; 90(2):347–55. https://doi.org/10.1016/j.ajhg.2011.12.010 PMID: 22284828
12. Alkuraya FS. Genetics and genomic medicine in Saudi Arabia. Molecular genetics & genomic medicine. 2014; 2(5):369–78. https://doi.org/10.1002/mgg3.97 PMID: 25333061
13. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nature Genetics. 2016; 48:1071–6. https://doi.org/10.1038/ng.3592 PMID: 27428751
14. Alsalem AB, Halees AS, Anazi S, Alshamekh S, Alkuraya FS. Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. PLoS Genet. 2013; 9(12):e1004030. https://doi.org/10.1371/journal.pgen.1004030 PMID: 24367280
15. Consortium GP. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245
16. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–91. https://doi.org/10.1038/nature19057 PMID: 27535533
17. Alsmadi O, Thareja G, Alkayal F, Rajagopalan R, John SE, Hebbar P, et al. Genetic substructure of Kuwaiti population reveals migration history. PloS one. 2013; 8(9):e74913. https://doi.org/10.1371/journal.pone.0074913 PMID: 24066156
18. Hunter-Zinck H, Musharoff S, Salit J, Al-Ali KA, Chouchane L, Gohar A, et al. Population genetic structure of the people of Qatar. The American Journal of Human Genetics. 2010; 87(1):17–25. https://doi.org/10.1016/j.ajhg.2010.05.018 PMID: 20579625
19. Rodriguez-Flores JL, Fakhro K, Agosto-Perez F, Ramstetter MD, Arbiza L, Vincent TL, et al. Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome research. 2016; 26(2):151–62. https://doi.org/10.1101/gr.191478.115 PMID: 26728717
20. Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, et al. Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. The American Journal of Human Genetics. 2003; 72(4):1058–64. https://doi.org/10.1086/374384 PMID: 12629598

21. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome research. 2008; 18(5):830–8. https://doi.org/10.1101/gr.7172008 PMID: 18385274

22. Loogväli E-L, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, Metspalu E, et al. Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Molecular Biology and Evolution. 2004; 21(11):2012–21. https://doi.org/10.1093/molbev/msh209 PMID: 15254257

23. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics. 2018; 19(4):220. https://doi.org/10.1038/nrg.2017.109 PMID: 29335644

24. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS genetics. 2011; 7(4):e1001373. https://doi.org/10.1371/journal.pgen.1001373 PMID: 21533020

25. Alkuraya F. Impact of new genomic tools on the practice of clinical genetics in consanguineous populations: the Saudi experience. Clinical genetics. 2013; 84(3):203–8. https://doi.org/10.1111/cge.12131 PMID: 23451714

26. Alkuraya FS. Discovery of mutations for Mendelian disorders. Human genetics. 2016; 135(6):615–23. https://doi.org/10.1007/s00439-016-1664-8 PMID: 27068822

27. Alazami AM, Hijazi H, Al-Dosari MS, Shaheen R, Hashem A, Aldahmesh MA, et al. Mutation in ADAT3, encoding adenosine deaminase acting on transfer RNA, causes intellectual disability and strabismus. Journal of medical genetics. 2013; 50(7):425–30. https://doi.org/10.1136/jmedgenet-2012-101378 PMID: 23620220

28. Abouelhoda M, Faquih T, El-Kalioby M, Alkuraya FS. Revisiting the morbid genome of Mendelian disorders. Genome biology. 2016; 17(1):235. https://doi.org/10.1186/s13059-016-1102-1 PMID: 27884173

29. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nature genetics. 2015; 47(11):1272. https://doi.org/10.1038/ng.3368 PMID: 26366554

30. Al-Owain M, Al-Zaidan H, Al-Hassnan Z. Map of autosomal recessive genetic disorders in Saudi Arabia: concepts and future directions. American Journal of Medical Genetics Part A. 2012; 158A(10):2629–40. https://doi.org/10.1002/ajmg.a.35551 PMID: 22903695

31. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010; 26(22):2867–73. https://doi.org/10.1093/bioinformatics/btq559 PMID: 20926424

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

33. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4(1):7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

34. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020; 367(6484). https://doi.org/10.1126/science.aay5012 PMID: 32193295

35. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016; 538(7624):201–6. https://doi.org/10.1038/nature18964 PMID: 27654912

36. Bradburd GS, Ralph PL, Coop GM. Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution. 2013; 67(11):3258–73. https://doi.org/10.1111/evo.12193 PMID: 24102455

37. Goudet J. Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes. 2005; 5(1):184–6.

38. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009; 19(9):1655–64. https://doi.org/10.1101/gr.094052.109 PMID: 19648217

39. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics. 2012; 8(11):e1002967. https://doi.org/10.1371/journal.pgen.1002967 PMID: 23166502

40. Milanesi M, Capomaccio S, Vajana E, Bomba L, Garcia JF, Ajmone-Marsan P, et al. BITE: an R package for biodiversity analyses. BioRxiv. 2017:181610.

41. Felsenstein J. PHYLIP-Phylogeny Inference Package (Version 3.2) Cladistics 5: 164–166. 1989.

42. Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013:genetics. 112.147330. https://doi.org/10.1534/genetics.112.147330 PMID: 23410830

43. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. The American Journal of Human Genetics. 2015; 97(3):404–18. https://doi.org/10.1016/j.ajhg.2015.07.012 PMID: 26299365

44. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics. 2018; 103(3):338–48. https://doi.org/10.1016/j.ajhg.2018.07.015 PMID: 30100085

45. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics. 2007; 81(5):1084–97. https://doi.org/10.1086/521987 PMID: 17924348

46. Zhou Y, Browning SR, Browning BL. A fast and simple method for detecting identity by descent segments in large-scale data. The American Journal of Human Genetics. 2020. https://doi.org/10.1016/j.ajhg.2020.02.010 PMID: 32169169

47. Jostins L, Xu Y, McCarthy S, Ayub Q, Durbin R, Barrett J, et al. YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data. arXiv preprint arXiv:14077988. 2014.

48. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids research. 2016; 44(W1):W58–W63. https://doi.org/10.1093/nar/gkw233 PMID: 27084951