BMC Genomic Data

RESEARCH

Open Access

Ancestry-related distribution of Runs of homozygosity and functional variants in Qatari population

Massimo Mezzavilla[1,2*], Massimiliano Cocca[1], Pierpaolo Maisano Delser[3], Ramin Badii[4], Fatemeh Abbaszadeh[4], Khalid Abdul Hadi[5], Girotto Giorgia[1,6] and Paolo Gasparini[1,6]

Abstract

Background: Describing how genetic history shapes the pattern of medically relevant variants could improve the understanding of how specific loci interact with each other and affect diseases and traits prevalence. The Qatari population is characterized by a complex history of admixture and substructure, and the study of its population genomic features would provide valuable insights into the genetic landscape of functional variants. Here, we analyzed the genomic variation of 186 newly-genotyped healthy individuals from the Qatari peninsula.

Results: We discovered an intricate genetic structure using ancestry related analyses. In particular, the presence of three different clusters, Cluster 1, Cluster 2 and Cluster 3 (with Near Eastern, South Asian and African ancestry, respectively), was detected with an additional fourth one (Cluster 4) with East Asian ancestry. These subpopulations show differences in the distribution of runs of homozygosity (ROH) and admixture events in the past, ranging from 40 to 5 generations ago. This complex genetic history led to a peculiar pattern of functional markers under positive selection, differentiated in shared signals and private signals. Interestingly we found several signatures of shared selection on SNPs in the *FADS2* gene, hinting at a possible common evolutionary link to dietary intake. Among the private signals, we found enrichment for markers associated with HDL and LDL for Cluster 1(Near Eastern ancestry) and Cluster 3 (South Asian ancestry) and height and blood traits for Cluster 2 (African ancestry).

The differences in genetic history among these populations also resulted in the different frequency distribution of putative loss of function variants. For example, homozygous carriers for rs2884737, a variant linked to an anticoagulant drug (warfarin) response, are mainly represented by individuals with predominant Bedouin ancestry (risk allele frequency G at 0.48).

Conclusions: We provided a detailed catalogue of the different ancestral pattern in the Qatari population highlighting differences and similarities in the distribution of selected variants and putative loss of functions. Finally, these results would provide useful guidance for assessing genetic risk factors linked to consanguinity and genetic ancestry.

Keywords: Qatar population, Admixture, Runs of homozygosity, Positive selection, Loss of function, Warfarin response

Background

Qatar has a rich and fascinating history, inhabited by humans for approximately 50,000 years with a substantial influx of Arab tribes from the surrounding region, mainly from the Nejd desert to the West. Islam began to flourish

*Correspondence: mzzvilla@gmail.com

[1] Institute for Maternal, and Child Health - IRCCS "Burlo Garofolo", Via dell'Istria 65/1, 34137 Trieste, Italy
Full list of author information is available at the end of the article

in Qatar in the seventh century CE, and the area became an important cultural centre for the spread of the Islamic religion [1]. Like many other Gulf region countries, the Qatari population is affected mainly by diabetes, obesity, and cardiovascular diseases [2], in particular The prevalence of obesity in Qatar is among the highest in the world, 41.4% based on reports from the Qatari Ministry of Public Health (https://phs.moph.gov.qa/data/healthy-lifestyle/), in addition the level of CVD related deaths in Qatar is high as in other high income countries" (https://phs.moph.gov.qa/data/cardiovascular-diseases/).

Thus, it is an interesting "laboratory" to investigate the genetics and environmental risk factors underlying such diseases.

As a matter of fact, genetic disorders are generally well-described by purifying selection models, while complex-disease susceptibility is tied, at least in part, to evolutionary adaptations and demography. In particular, reducing effective population size due to inbreeding and bottlenecks reduces the effectiveness of both positive and purifying selection [3]. The type of selection and the strength of its coefficient vary across populations, affecting the prevalence of causative variants for diseases and traits [4].

Previous data on the Qatari population demonstrated a peculiar clustering and different variance in homozygosity regions (ROH) [5]. Recent data show that ROH across genomes could impact different phenotype distributions across different ancestries [6, 7]. Such changes in the genomic architecture of a given population could also impact the effect of the same variants in different populations. For example, although PPARγ gene variants are associated with diabetes in some individuals of European descent, mutations in this gene were found not to be a risk factor in the Qatari population [8].

In addition, a recent study showed that European-derived polygenic scores (PGS) had reduced predictive performance in the Qatari population [9].

In fact, several studies investigated the pattern of genetic diseases in conjunction with endogamy and consanguinity in the populations of this geographical area [10–12].

An essential piece of information needed is the knowledge of the genetic history and the evolutionary mechanism behind the genomic makeup of the Qatar population. A recent work studied several thousands of individuals highlighted the link with ancient hunter-gatherers and Neolithic farmers from the Levant [13]. However, in our work we aimed to integrate several pieces of information coming from population genetics analyses and we tried to integrate them in order to understand the pattern of deleterious variation in a group of Qatari individuals.

Here, we investigated the genetic structure of 186 newly genotyped individuals from Qatar and analyzed the distribution of ROH regions under recent natural selection and putative loss of function variants.

Our work aims to address the following questions: i) How genetic structure and demography affect the ROH pattern in the Qatari population and ii) how genetic structure affects the pattern of genes under putative positive selection and the distribution of deleterious variants with a specific focus on the loss of function variants. Our final goal is to provide a detailed insight into the genetic makeup of the Qatari population to better estimate and understand the genetic risk factors based on ancestry components, demography and natural selection.

Results

Uniparental markers analysis

High variation was observed for mitochondrial DNA (haplotype diversity = 0.873) in both the entire dataset and the subset, including male individuals only. Major haplogroups are represented by H (South West Asia origin), L (Africa origin) and J (Western Asia origin). The Y chromosome shows a reduced diversity with a major haplogroup (J1*) representing 75% of the Y chromosomes analysed (see Fig. S1 A-B-C). The ratio of Y chromosome haplotype diversity (haplotype diversity = 0.574) on mitochondrial haplotype diversity is 0.65.

Population structure and admixture

An unsupervised analysis with ADMIXTURE v.1.3 [14] was performed on the Qatari samples using a subset of reference population from the Human Origins dataset downloaded from https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data and the lowest cross-validation error was obtained with a total number of cluster equal to 11 (see Table S1. Four major ancestral components differently distributed among individuals were detected (Fig. 1A). The red component was found mainly in the Bedouin population. The green component (found mainly in Yoruba samples) was appreciable only in a fraction of the Qatari sample. The violet (Palestinian) and azure (South Asian) components were found in another group of Qatari individuals showing low levels of both red (Bedouin) and green (African) components. A small group of individuals shows an admixture pattern that contains only South Asian and East Asian ancestry but neither Palestinian nor Bedouin. A full representation of all cluster solution is shown in Fig. S2.

Using the first six principal components from the Principal component analyses (PCA), a gaussian clustering using the approach implemented in Mclust [15] was carried out. An overall number of four clusters was detected:

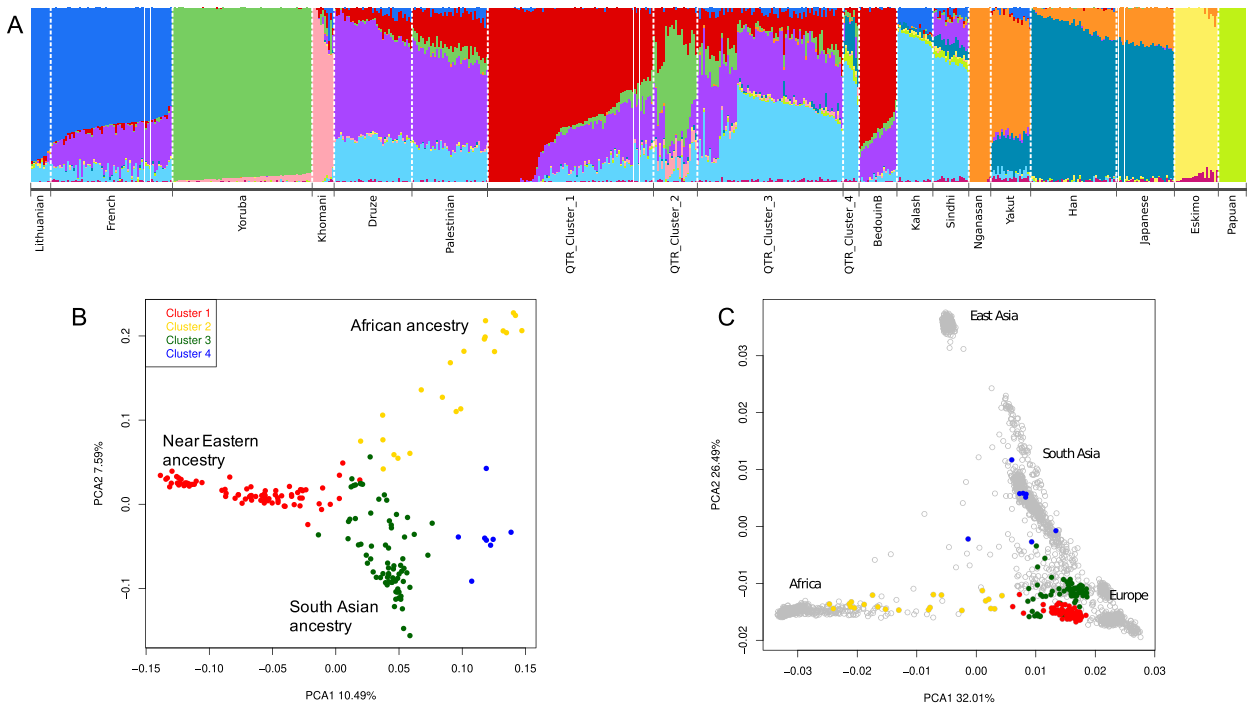

Fig. 1 Population structure of Qatar (**A**) Admixture plot for K = 11 using Human Origins dataset. The red colour represents the Middle Eastern-Bedouin like ancestry, the cyan colour represents the South Asian component the violet component represent the Middle Eastern-Palestinian like component, the green one represents the African component and the blue one represents the East Asian component (**B**) Principal component analysis followed by Gaussian clustering based only on 186 Qatari sample, a total of four clusters were found. **C** Projected principal component of 186 Qatari sample onto 1000G populations. We can observe how Cluster 2 shows African ancestry, Cluster 3 has South Asian Ancestry and Middle Eastern Ancestry, Cluster 1 has the highest Bedouin like ancestry and finally Cluster 4 shows evidence of South Asian and east Asian ancestry.

Cluster 1 (red), Cluster 2 (gold), Cluster 3 (green) and finally Cluster 4 (blue), which contains a small fraction of East Asian ancestry (Fig. 1B, Figs. S3-S4).

PCA using as reference the 1000 Genome Project [16] data shows that the Qatari individuals are placed between the European and South Asian pole of variation. Cluster 2 (gold) is spread towards African samples while individuals from Cluster 4 show similar variation to East Asian samples, confirming the ADMIXTURE analysis (Fig. 1C).

In order to better investigate the genetic relationships between individuals using admixture patterns, we used the individual ancestry values obtained from previous admixture analyses to build a distance matrix which was used to generate a dendrogram (Fig. 2A). Each individual was coloured according to their cluster assignment, and for each of them, the level of homozygosity due to ROH (Runs of homozygosity) was collected. As shown in Fig. 2B**,** individuals from Cluster 1 and Cluster 3 show the highest level of ROH. These clusters are characterized by Bedouin and Palestinian/South Asian ancestry. On the other hand, individuals from Cluster 2 (characterized by the highest level of African ancestry) shows the lowest level of runs of homozygosity in our dataset. Interestingly, individuals from Cluster 4 (with both South Asian and East Asian ancestry) show a homozygosity level similar to that of Cluster 3.

We also found a sizeable ancestry-related variation in the number of ROH segments and total homozygosity due to ROH when the three clusters were compared to the 1000 Genome populations (Fig. 3A, Fig. S5). We should note that Cluster 1 and Cluster 3 have increased total homozygosity with respect to the average number of segments, which suggests recent consanguinity [17, 18]

Such a diverse distribution could be explained by the different genetic history of each cluster.

Analyses of effective population size (Ne) in the last 50 generations using IBDNe [19] further support a significant difference in the level of genetic drift, as the confidence intervals of the effective population sizes across generations never overlap between each other (Fig. 3B).

Admixture analysis using MALDER [20] revealed several admixture events that happened at different times: one admixture event between 32 ± 3 generations ago for Cluster 1 (in which the reference populations with the

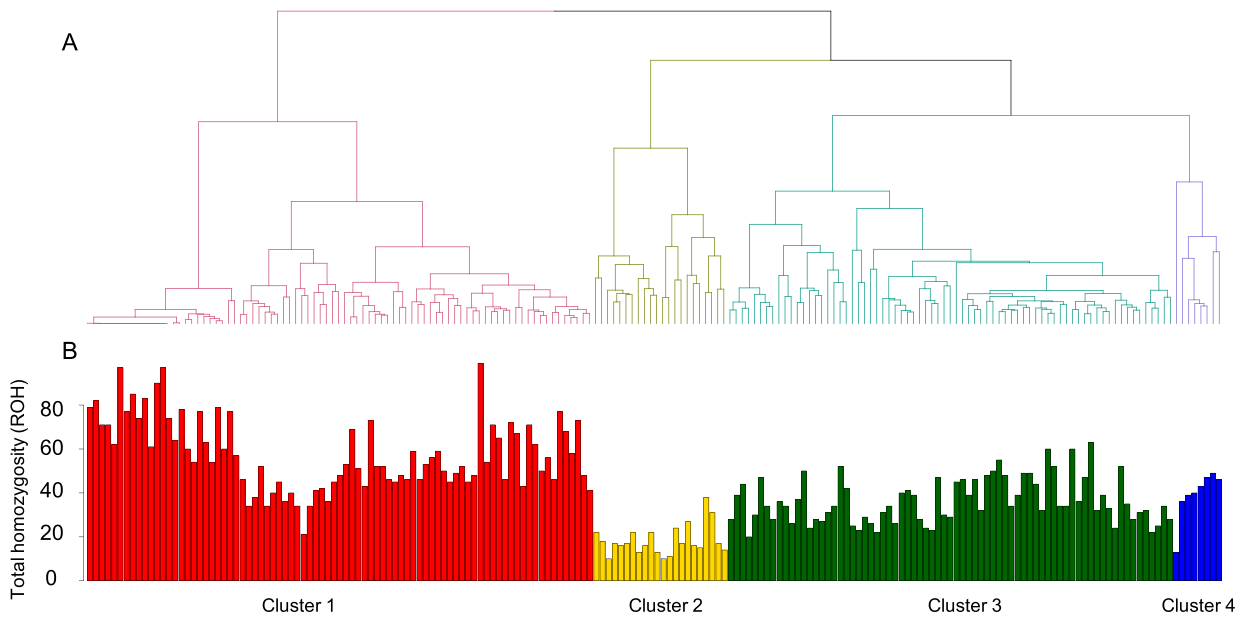

Fig. 2 Clustering based on admixture Q values and ROH pattern. **A** dendrogram clustering and **B** each bar represents the individual total homozygosity due to ROH in Mb. Cluster 1 is colored in red, Cluster 2 in yellow, Cluster 4 in dark green and Cluster 4 in blue

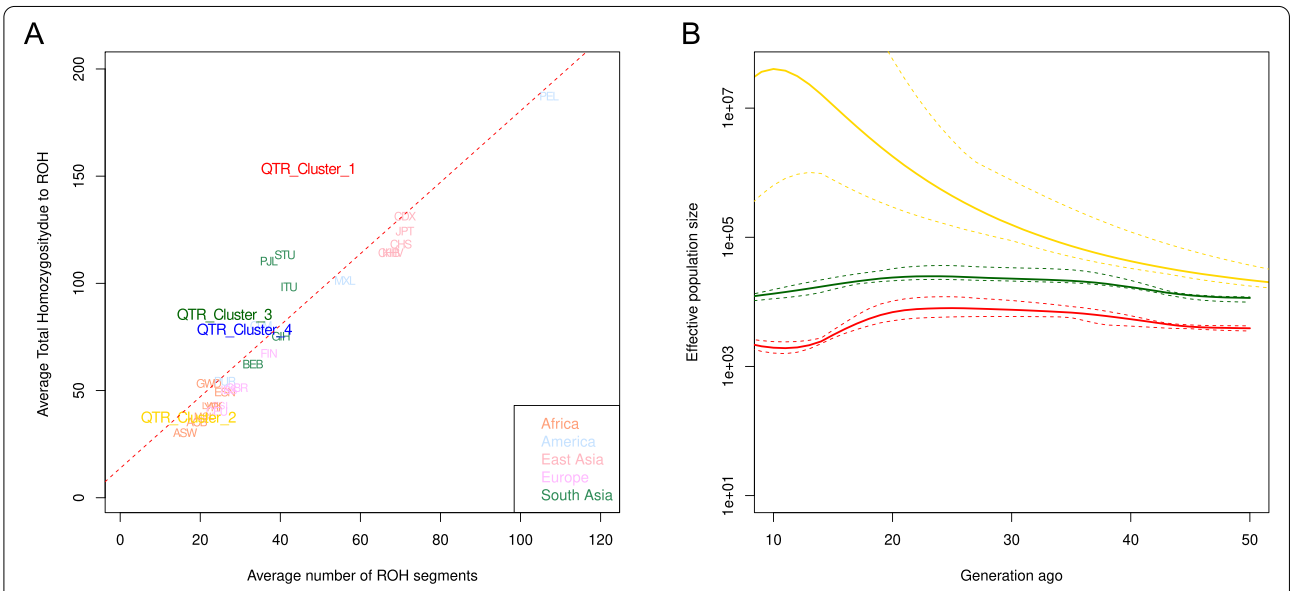

Fig. 3 **A** Average level of total homozygosity and number of ROH segments in Qatari population. On the y-axis the ROH were measured in Mb. Population from the 1000G are coloured according to their geographic origin. **B** Effective population size (Ne) estimates with IBDNe. We estimated the Ne of the three major clusters found in the Qatari samples. Dotted lines represent the 95% CI

highest Zscore, according to MALDER are Greek and Yoruba) and a more recent event 5 ± 0.5 generations for Cluster 2 (reference populations with the highest Z score, according to MALDER are Biaka and Greek).

Interestingly, Cluster 3 shows evidence of two admixture events: one at 42 ± 5 generations ago (reference populations: Greek and Yoruba) and one more recently at 5 ± 1 generation ago (reference populations:

Biaka and French). We had to exclude Cluster 4 from this analysis as its small sample size could produce unreliable results in detecting admixture events.

Selection signals

We applied the NSL statistic [21] to the three major clusters found in our dataset to understand how genetic structure affects the pattern of genes under putative positive selection. A conservative approach was considered, collecting only the results of markers previously associated with a phenotype, using, as a reference, the GWAS catalogue.

Scans for selection signals revealed that most hits are private to each cluster if we consider all the signals putatively functional (NSL > = 99th percentile of the genomic distribution and presence in GWAS catalogue (see Figs. S6-S7).

Among the top signals that are shared between all three major clusters (NSL score over the 99th percentile and SNP present in GWAS catalogue), we found two variants in *the FADS2* gene: one rs174578, rs174583 and rs174601 associated with haemoglobin, serum metabolite measurement and different lipid traits (HDL, LDL), respectively [22–24]. Additional signals of shared selection signatures were found in *RYR1,* where the variant rs3786829 was associated with peanut allergy [25], another SNP was found in *DENND1A,* the variant carrying the signal (rs2479106) was associated with polycystic ovary syndrome [26]. Finally, we found selection signatures in three additional markers, one associated with microglial activation measurement (rs651691) [27], one associated response to anti-depressant treatment in major depressive disorder (rs10517287) [28] and one associated with trans-fatty acid levels (rs17099388) [29] (Table 1). Then we grouped the private signals of selection accordingly to the associated phenotypes. We discovered signatures of selection for genes linked to lipid traits, BMI and serum metabolite levels for Cluster 1 and Cluster 3. For Cluster 2, we found signals in SNPs involved in blood traits and height (Table S2).

Putative loss of function variation

Finally, we investigated how genetic structure affected the distribution and prevalence of loss of function variants. A total of 97 putative loss of function variants (LOF) were analyzed using a custom-made list described in the Method section. For thirty of them, a significant difference in frequency was found (after Bonferroni correction) only in one cluster compared to the others (see Table S3). The majority of them are specific to Cluster 1 (which shows higher homozygosity and Bedouin-like ancestry) and Cluster 2 (African ancestry). The markers with the highest difference in frequency in each cluster were then further analyzed (top five lowest p-values, corresponding to the top 2% of the results). One of them, rs2884737 (p-value = 5E-07), is located within *the VKORC1* gene and detected at high frequency in Cluster 1(Near Eastern ancestry). This variant is involved in warfarin response [30], A graphical representation of how ancestry determined the genotype distribution of these variants is shown in Figs. S6, S7 and S8. In Cluster 2(African ancestry), signals for rs1127745 located in *ACOX2* and associated with triglyceride levels [31]. One variant, rs35400274 (in *C17orf107*, a gene associated with Sphingomyelin levels [32], was present in Cluster 3 (South Asian ancestry). Finally, one variant, rs3213755, in *the KRTAP1-1* gene, which encodes for a keratin-associated protein, was found in Cluster 4; to our knowledge, there are no phenotypes previously associated with this gene. To investigate the relationship between effective population size and LOF distribution we applied the following approach: we grouped the LOF variants into two groups. The first one comprises high deleteriousness variants using CADD score [33] as measure

Table 1 Shared signals of selection among the different subgroups in Qatar

SNP_id	Location	Consequence	SYMBOL	Gene	NSL Cluster1	NSL cluster2	NSL cluster3
rs651691	1:193,958,320–193,958,320	intergenic_variant	-	-	-3.42	-2.86	-2.66
rs10517287	4:33,624,702–33,624,702	intergenic_variant	-	-	-3.73	-3.12	-3.51
rs17099388	5:142,095,250–142,095,250	intergenic_variant	-	-	-3.36	-3.81	-3.15
rs2479106	9:126,525,212–126,525,212	intron_variant	DENND1A	ENSG00000119522	-2.83	-2.72	-3.04
rs174577	11:61,604,814–61,604,814	intron_variant	FADS2	ENSG00000134824	-2.95	-2.79	-3.14
rs174578	11:61,605,499–61,605,499	intron_variant	FADS2	ENSG00000134824	-3.63	-3.04	-3.19
rs174601	11:61,623,140–61,623,140	intron_variant	FADS2	ENSG00000134824	-3.96	-3.11	-3.63
rs3786829	19:39,014,184–39,014,184	intron_variant	RYR1	ENSG00000196218	-2.82	-3.79	-3.45

of deleteriousness (CADD > = 25), and the second one including low deleteriousness ones (CADD < = 5); then, we estimated the median allele frequency in each group and each cluster found in the Qatari sample. The amount of low deleteriousness variation is related to the level of drift, and the amount of high deleteriousness variation indicates the natural selection efficiency. The ratio of high deleteriousness variation to low deleteriousness variation should hint at the efficiency of selection. A low ratio indicates higher purifying selection efficiency compared to drift. A high ratio suggests that selection is less efficient compared to genetic drift. As we can observe from Table S5 the lowest ratio is from Cluster 2 and the highest is from Cluster 1 which indicates that in the population with highest Ne, selection is more efficient.

Discussion

Previously published works [5, 9, 34–36] described the different ancestral components in the Qatari population. Our focus is to describe how a peculiar genetic history shaped one population's genomic pattern in terms of homozygosity burden, variants under positive selection, and genetic drift of putative loss of function variants. With the current emphasis on precise and personalized medicine, and therefore on rare variants, we must not forget that demography and admixture shape the prevalence of common genetic factors that could impact the phenotype distribution at a population level, with repercussion on the welfare system.

With our findings, we provide a more comprehensive analysis regarding the ancestry-related structure that could be useful for future analyses on both array and whole-genome sequencing data (WGS). Three major ancestral groups (with predominantly Bedouin, African, and South Asian ancestry) named Cluster 1, Cluster 2 and Cluster 3 were found in agreement with previous data and uniparental marker analysis. The difference in variability between Y and mitochondrial data could hint at a sex-biased migration, in fact an higher haplotype variability in the mitochondrial genome respect to the Y chromosome could hint to movement of females in patrilocal groups [37]. Interestingly, a novel cluster with a small fraction of East Asian ancestry was found (Cluster 4), indicating additional cryptic gene flow from a more distant origin in the past. This additional cluster suggests that increased sample size could reveal higher levels of substructure than expected, further hinting at the Qatari population as a melting pot of different ancestries and admixture events [13]. Moreover, this scenario adds a new layer of complexity to the genetic architecture of the Qatari population. Therefore, for example, GWAS analysis should carefully consider this complex stratification to avoid any bias, for example, performing association studies in each ancestral subgroup separately, if possible, or selecting a method that can correctly take into account the cryptic structure of this and similar populations [38–40].

Our data showed how the population substructure is linked to the difference in ROH pattern, which affects phenotype distribution [6, 7, 41]. Cluster 1 showed higher levels of ROHs with respect to Cluster 2, Cluster 3 and Cluster 4, consequently. Overall, the present findings suggest a hierarchical level of population substructure in the Qatari population, characterized by varying levels of homozygosity. One limitation of our study is the lack of phenotype information. Despite some variants are found in homozygous state in a population, it is difficult to predict the overall variability of a phenotype linked to these markers, mainly because the majority of associated genetic variants explain very little of the phenotype variance.

Additional analyses revealed a different effective population size (Ne) between the three major clusters in recent time, such as the timing and number of admixture events. If we consider a generation time of 30 years, the time of the admixture events for cluster 1 is around 32 generations ago ~ 1040 CE (32 generations) while for Cluster 2 is ~ 1859 CE (5 generations). Cluster 3 shows two admixture events, one at 1859 CE (similar to cluster 2) and one at ~ 740 CE (42 generations ago). It is interesting how we can roughly overlap the admixture events for Cluster 1 and Cluster 3 to the period of the Abbasid Caliphate (750–1258 CE), where the Qatari region started to become a strategic economic hub, and pearl trading flourished. The most recent admixture events (for Cluster 3 and Cluster 2) correspond to the first stage in Qatar's development as a sheikhdom in recent history when the house of Thani started to rise in power [42]. Cluster 1 is the genetic group with lowest effective population and no evidence of recent admixture.

These results lead us to the assumption that also, the role of natural selection could be different. For this reason, we investigated the pattern of recent selection using nSL statistics. The analysis revealed that, despite all clusters sharing the same environment and actual geographical location, the selection signals are composed predominantly of private ones (~ 70%). These signals involve markers previously associated with lipid traits such as HDL and LDL (Cluster 1 and Cluster 3) and height and blood traits (Cluster 2).

Some of the signals are shared between clusters, such as variants in *FADS2*, which could be linked to diet adaptation [43, 44]. The pattern of shared signals is negatively correlated with the genetic distance between these three clusters. As previously shown, the selection pressure should come from an adaptation to a diet characterized

by a high level of fatty acids derived from plants but relatively poor in fatty acids derived from fish or mammals [45] which could relate to the introduction of agriculture in the Middle East. One limitation in our analysis is that we based our assumptions on selection taking into account only specific variants reported in GWAS catalogue. Considering only a direct effect on a trait could restrict the possible explanations of selection pressures.

Besides signals of selection (related to ancestral origins), genetic drift shows different patterns in the Qatari population. Due to the reduced effective population size, we also expect reduced effectiveness of purifying selection. Thus, we investigated the pattern of a specific group of variants: the putative loss of function variants (pLOF). Our analysis revealed that there is a relative higher ratio of deleterious LOF (CADD > = 25) in the clusters with lower Ne (Cluster1and Cluster 3), respect to the Cluster 2(Africans), which shows higher effective population size.

Our work showed that several common putative pLOF harbour significant differences in allele frequency between clusters. Some of them, like the variant in the VKORC*1* gene, are linked to a specific pharmacological response and show higher prevalence in Cluster 1 or are considered risk factors for phenotypes like triglyceride level (*ACOX2* variant for Cluster 2).

The result on *VKORC1* is of particular interests, mainly because recent works showed the importance of warfarin management in the Qatar population [46] and how this gene is involved in warfarin dose variability in Qatari [47].Here we show that one genotype is more prevalent in one ancestry respect to another in the structured population of Qatar.

A study of the population structure of Qatar's people, as inferred by genetic testing, is necessary to determine how best to perform several association studies and other genetically-assisted analyses of risk in the Qatari population. Furthermore, our findings provide crucial information for risk stratification in the Qatari population.

Material and methods

Data preparation

Saliva samples from 188 healthy individuals were collected in Hamad Medical Corporation (HMC), A written informed consent for participation was obtained from all subjects. Samples DNA was extracted at the IRCCS Burlo Garofolo Hospital. Genotyping was conducted at the Life & Brain Research Centre (Bonn, Germany) using the Illumina Infinium Global Screening Array-24 v1.0 (GSAMD-24v1-0_20011747_A1). The initial quality control was performed on Illumina GenomeStudio software to remove poorly called samples and sites. Raw genotype data underwent a step of recalling using the software z-call [48] to obtain more reliable calls on low-frequency variants. PLINK v1.9 software [49] was used to process the genotype calls for further variants and samples QC: i) remove samples with high IBD sharing; ii) remove sites with a heterozygous rate higher than three standard deviations from the mean heterozygosity rate distribution; iii) remove sites and samples by call rate (–geno 0.01 –mind 0.05 options); iv) remove sites, not in Hardy–Weinberg equilibrium (–hwe 0.000001 option). The dataset resulting from these QC steps resulted in 186 individuals that was finally phased using the shapeit2 software [50], without using any reference panel.

Y and mitochondrial haplogroup analysis

First, 28 male samples were extracted from the dataset and Y chromosome haplogroups were assigned using AMY-tree v2.0 software [51]. Input files were created by converting PED and FAM files into a vcf using PGDSpider v2.1.1.1 [52] and then from a vcf into AMY-tree input files with R scripts. Results were then combined using in-house R scripts. Mitochondrial analysis of 186 individuals was performed using the software haplogrep-2.1.20 [53]. Haplotype diversity was estimated following the formula described in [54]

Population structure and admixture pattern

To obtain a larger picture of the geographical pattern we merged our dataset with 1000G Phase 3 [16] (dataset-A) and Human Origins dataset [55] (dataset-B). Principal component analyses on dataset-A and dataset-B were performed after removing markers in linkage disequilibrium using the option –indep-pairwise 200 50 0.4 implemented in PLINK [49]. Clustering approach was made using the R package Mclust [15] on the first 6 PCA eigenvectors. A complete list of all population used and their relative sample size is reported in Table S4.

Unsupervised admixture analysis using ADMIXTURE v1.23 [14] was done on dataset-B after removing the populations with less than ten individuals. Time of admixture using all possible combinations of reference populations was performed using MALDER [20].

Inbreeding and runs of homozygosity estimates were calculated using PLINK using the option –homozyg and –het.

We further investigated effective population size using IBDseq [56] and IBDNe [19] on each genetic cluster identified. We using a threshold of 2 centimorgan for IBD segments and default parameters as suggested for SNP array data.

Selection scan

Selection scans in the different subgroups were done using the nSL statistic, a modification of iHS that has improved power in detecting soft sweeps [21]. Genotype

data were phased using Eagle [57], and nSL statistics were estimated and normalized using selscan [58]. First, we collected the values with a score over than 2 and present in GWAS catalogue reported the fraction of private and shared variants under putative positive selection between the various subgroups. We then collected the results falling over the 99.9th percentile of the distribution of genomic nSL and we selected the variants reported in the GWAS catalogue. These analyses were done in order to assess the impact of natural selection in putatively functional variants already associated with disease or traits.

Putative loss of function variant distribution

We created a manually curated dataset of LOF variants which was composed by two lists: the first set was a list of loss of function variants described in MacArthur et al. [59] while the second list was composed by all variants annotated as stop-gain using VEP tool [60]. This selection aimed to obtain a reliable list of putative loss of function variants. We grouped the LOF into two categories: one with CADD score > = 25 which are considered as high deleterious and one with CADD score < = 5, which are considered as low deleterious. We estimated the average allele frequency in each group in each genetic cluster.

For each LOF, using the function –assoc implemented in PLINK, we selected the differentiated ones on one cluster but not in the other. Only variants showing significant p-values after Bonferroni correction were further analyzed.

We investigated how ancestry affects the distribution of genotypes using the R package party [61], selecting the top differentiated markers in each subpopulation.

Abbreviations

PCA: Principal component analysis; ROH: Runs of homozygosity; Ne: Effective population size; GWAS: Genome wide association study; nSL: Number of segregating sites by length; HDL: High-density lipoprotein; LDL: Low-density lipoprotein; BMI: Body mass index; pLOF: Putative loss of function; IBD: Identity by descent; QC: Quality control.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-022-01087-1.

Additional file 1: Figure S1. Distribution of Y and mitochondrial haplogroups in Qatari samples. **Figure S2.** Admixture runs from K=3 to K=11.At K=3 we can observe the separation between the Cluster 2 (with African ancestry) and the other three clusters, and we can observe the higher East Asian ancestry in Cluster 4. From K=5 to K=7 we can observe how the Bedouin ancestry is predominant in the Cluster 1 respect to the other clusters. At K=9 we can see how the South Asian (Sindhi) ancestry is becoming predominant in the Cluster 3. **Figure S3.** Principal component analysis.PC3 versus PC4 on the left panel, PC5 versus PC6 on the right panel. Variance explained by each axis is reported as well. **Figure S4.** Ancestry proportions in the Qatari sample. The red colour represents the Middle Eastern-Bedouin like ancestry, the cyan colour represents the South Asian component the violet component represents the Middle Eastern-Palestinian like component, the green one represents the African component and the blue one represents the East Asian component. **Figure S5.** Beanplot of Total Homozygosity in Qatar and 1000G populations. The dotted line represents the average worldwide level of total homozygosity. **Figure S6.** Genomic distribution of unstandardized —nSL— score. Each line represents the distribution of —nSL— values in each cluster, the dotted line represents the cut-off to discriminate between putatively under selection and neutral markers. **Figure S7.** Venn Diagram of shared and private signal of selection. The numbers represent the number of SNPs with nSL score over the 99th percentile of the genomic distribution and previously associated to a phenotype (GWAS catalogue) are reported. We can observe how the majority of the variants under selection and previously associated with a phenotype are private of each cluster and only a small fraction is shared between all of them. **Figure S8.** Regression tree analyses of rs2884737. The analysis shows the different and significant genotype distribution of rs2884737 among the four different cluster found in our dataset. The C allele in homozygous state is more prevalent in Cluster 1. **Figure S9.** Regression tree analyses of rs1127745. The analysis shows the different and significant genotype distribution of rs1127745 among the four different cluster found in our dataset. The G allele in homozygous state is more prevalent in the Cluster 2. **Figure S10.** Regression tree analyses of rs35400274. The analysis shows the different and significant genotype distribution of rs35400274 among the four different cluster found in our dataset. The A allele in homozygous state is more prevalent in the Cluster 1 and Cluster 2.

Additional file 2: Table S1. Cross validation errors from admixture runs.

Additional file 3: Supplementary Table 1. Phenotypes associated with private Signals of Selection among the different subgroups in Qatar.

Additional file 4: Supplementary Table 2. Signficantly differentisted pLOF variants between Qatari genetic clusters.

Additional file 5: Supplementary Table 3. Populations with relative sample size used in this study.

Additional file 6: Table S5. Level of diversity of low deleterious and high deleterious variants.

Acknowledgements

We would like to thank all the participants in the study

Authors' contributions

Sample collection was performed by RB, FA, KAH. Funding acquisition from RB, PG, FA, GG, KAH. MM performed the study design. MM, PMD, MC performed data analysis. MM, PMD, MC, GG and PG wrote the manuscript. All authors read and approved the manuscript.

Funding

This work was supported by Qatar National Research Fund, Age-related hearing loss in Qatar: a genomic approach to identify causative gene (JSREP07-013–3-006), and this work was also supported by IRCCS Burlo Garofolo of Trieste (5X1000 funding).

Availability of data and materials

A vcf file including all the variants with information on allele frequencies in the whole dataset, has been submitted to the European Variation Archive (EVA), study accession number: PRJEB51505. The data is accessible at the following link: https://www.ebi.ac.uk/ena/data/view/PRJEB51505**.** Data are also available upon request from the authors.

Declarations

Ethics approval and consent to participate

The study was conducted according to the Declaration of Helsinki guidelines and approved by the Ethics Committee of the Institute for Maternal and Child Health—I.R.C.C.S. "Burlo Garofolo" of Trieste (Italy) (2007 242/07). A written informed consent for participation was obtained from all subjects.

Consent for publication

Not applicable.

Competing interests
On behalf of all authors, the corresponding author states that they have no competing interests.

Author details
[1]Institute for Maternal, and Child Health - IRCCS "Burlo Garofolo", Via dell'Istria 65/1, 34137 Trieste, Italy. [2]Department of Biology, University of Padua, Padua, Italy. [3]Department of Zoology, University of Cambridge, Cambridge, England. [4]Molecular Genetics Laboratory, Laboratory of Medicine and Pathology, Hamad Medical Corporation (HMC), Doha, Qatar. [5]Audiology and Balance Unit, National Program for Early Detection of Hearing Loss, Hamad Medical Corporation (HMC), Doha, WH, Qatar. [6]Department of Surgical, Medical and Health Sciences, University of Trieste, 34149 Trieste, Italy.

Received: 21 February 2022 Accepted: 29 August 2022
Published online: 21 September 2022

References

1. Michalopoulos S, Naghavi A, Prarolo G. Trade and geography in the spread of Islam. Econ J. 2018;128:3210–41.
2. Turk-Adawi K, Sarrafzadegan N, Fadhil I, Taubert K, Sadeghi M, Wenger NK, et al. Cardiovascular disease in the Eastern Mediterranean region: epidemiology and risk factor burden. Nat Rev Cardiol. 2018;15:106–19.
3. Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nat Commun. 2017;8:1–7.
4. Prohaska A, Racimo F, Schork AJ, Sikora M, Stern AJ, Ilardo M, et al. Human disease variation in the light of population genomics. Cell. 2019;177:115–31.
5. Hunter-Zinck H, Musharoff S, Salit J, Al-Ali KA, Chouchane L, Gohar A, et al. Population genetic structure of the people of Qatar. Am J Hum Genet. 2010;87:17–25.
6. Johnson EC, Evans LM, Keller MC. Relationships between estimated autozygosity and complex traits in the UK Biobank. PLoS Genet. 2018;14:e1007556.
7. Clark DW, Okada Y, Moore KHS, Mason D, Pirastu N, Gandin I, et al. Associations of autozygosity with a broad range of human phenotypes. Nat Commun. 2019;10(1):4957.
8. O'Beirne SL, Salit J, Rodriguez-Flores JL, Staudt MR, Abi Khalil C, Fakhro KA, et al. Exome sequencing-based identification of novel type 2 diabetes risk allele loci in the Qatari population. PLoS ONE. 2018;13:e0199837.
9. Thareja G, Al-Sarraj Y, Belkadi A, Almotawa M, Suhre K, Albagha OME. Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. Nat Commun. 2021;12:1–10.
10. Rodriguez-Flores JL, Fakhro K, Hackett NR, Salit J, Fuller J, Agosto-Perez F, et al. Exome sequencing identifies potential risk variants for Mendelian disorders at high prevalence in Qatar. Hum Mutat. 2014;35:105–16.
11. Al-Gazali L, Hamamy H, Al-Arrayad S. Genetic disorders in the Arab world. BMJ. 2006;333:831–4.
12. Fakhro KA, Robay A, Rodrigues-Flores JL, Mezey JG, Al-Shakaki AA, Chidiac O, et al. Point of care exome sequencing reveals allelic and phenotypic heterogeneity underlying Mendelian disease in Qatar. Hum Mol Genet. 2019;28:3970–81.
13. Razali RM, Rodriguez-Flores J, Ghorbani M, Naeem H, Aamer W, Aliyev E, et al. Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. Nat Commun. 2021;12:1–16.
14. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
15. Fraley C, Raftery AE. MCLUST version 3: an R package for normal mixture modeling and model-based clustering. DTIC Document; 2006.
16. Consortium GP. A global reference for human genetic variation. Nature. 2015;526:68.
17. McQuillan R, Leutenegger A-L, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008;83:359–72.
18. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS ONE. 2010;5:e13996.
19. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet. 2015;97:404–18.
20. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci. 2014;111:2632–7.
21. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol. 2014;31:1275–91.
22. Chen M-H, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell. 2020;182:1198–213.
23. Gallois A, Mefford J, Ko A, Vaysse A, Julienne H, Ala-Korpela M, et al. A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. Nat Commun. 2019;10:1–13.
24. Tintle NL, Pottala JV, Lacey S, Ramachandran V, Westra J, Rogers A, et al. A genome-wide association study of saturated, mono-and polyunsaturated red blood cell fatty acids in the Framingham Heart Offspring Study. Prostaglandins Leukot Essent Fatty Acids. 2015;94:65–72.
25. Liu X, Hong X, Tsai H-J, Mestan KK, Shi M, Kefi A, et al. Genome-wide association study of maternal genetic effects and parent-of-origin effects on food allergy. Medicine. 2018;97(9):e0043.
26. Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, et al. Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. Nat Genet. 2012;44:1020–5.
27. Felsky D, Roostaei T, Nho K, Risacher SL, Bradshaw EM, Petyuk V, et al. Neuropathological correlates and genetic architecture of microglial activation in elderly human brain. Nat Commun. 2019;10:1–12.
28. Clark SL, Adkins DE, Aberg K, Hettema JM, McClay JL, Souza RP, et al. Pharmacogenomic study of side-effects for antidepressant treatment options in STAR* D. Psychol Med. 2012;42:1151–62.
29. Mozaffarian D, Kabagambe EK, Johnson CO, Lemaitre RN, Manichaikul A, Sun Q, et al. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. Am J Clin Nutr. 2015;101:398–406.
30. Li T, Lange LA, Li X, Susswein L, Bryant B, Malone R, et al. Polymorphisms in the VKORC1 gene are strongly associated with warfarin dosage requirements in patients receiving anticoagulation. J Med Genet. 2006;43:740–4.
31. Johansson Å, Curran JE, Johnson MP, Freed KA, Fenstad MH, Bjørge L, et al. Identification of ACOX2 as a shared genetic risk factor for preeclampsia and cardiovascular disease. Eur J Hum Genet. 2011;19:796–800.
32. Tabassum R, Rämö JT, Ripatti P, Koskela JT, Kurki M, Karjalainen J, et al. Genetic architecture of human plasma lipidome and its link to cardiovascular disease. Nat Commun. 2019;10:1–14.
33. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310.
34. Mbarek H, Devadoss Gandhi G, Selvaraj S, Al-Muftah W, Badji R, Al-Sarraj Y, et al. Qatar genome: Insights on genomics from the Middle East. Hum Mutat. 2022;43(4):499–510.
35. Omberg L, Salit J, Hackett N, Fuller J, Matthew R, Chouchane L, et al. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. BMC Genet. 2012;13:49.
36. Mezzavilla M, Vozzi D, Badii R, Khalifa Alkowari M, Abdulhadi K, Girotto G, et al. Increased rate of deleterious variants in long runs of homozygosity of an inbred population from Qatar. Hum Hered. 2015;79:14–9.
37. Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, et al. Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. PLoS Genet. 2006;2:e53.
38. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50:1335–41.
39. Jørsboe E, Albrechtsen A. Efficient approaches for large-scale GWAS with genotype uncertainty. G3. 2022;12:jkab385.
40. Sesia M, Bates S, Candès E, Marchini J, Sabatti C. False discovery rate control in genome-wide association studies with population structure. Proc Natl Acad Sci. 2021;118:e2105841118.

41. Mezzavilla M, Navarra CO, Di Lenarda R, Gasparini P, Bevilacqua L, Robino A. Runs of homozygosity are associated with staging of periodontitis in isolated populations. Hum Mol Genet. 2021;30:1154–9.
42. Rahman H. The emergence of Qatar: the turbulent years, 1627–1916. Routledge; 2005.
43. Maisano Delser P, Ravnik-Glavač M, Gasparini P, Glavač D, Mezzavilla M. Genetic landscape of Slovenians: past admixture and natural selection pattern. Front Genet. 2018;9:551.
44. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528:499–503.
45. Buckley MT, Racimo F, Allentoft ME, Jensen MK, Jonsson A, Huang H, et al. Selection in Europeans on fatty acid desaturases associated with dietary changes. Mol Biol Evol. 2017;34:1307–18.
46. Elewa H, Alhaddad A, Al-Rawi S, Nounou A, Mahmoud H, Singh R. Trends in oral anticoagulant use in Qatar: a 5-year experience. J Thromb Thrombolysis. 2017;43:411–6.
47. Bader L, Mahfouz A, Kasem M, Mohammed S, Alsaadi S, Abdelsamad O, et al. The effect of genetic and nongenetic factors on warfarin dose variability in Qatari population. Pharmacogenomics J. 2020;20:277–84.
48. Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, et al. zCall: a rare variant caller for array-based genotyping Genetics and population analysis. Bioinformatics. 2012;28:2543–5.
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007;81:559–75.
50. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10:e1004234.
51. Van Geystelen A, Decorte R, Larmuseau MHD. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics. 2013;14:1–12.
52. Lischer HEL, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics. 2012;28:298–9.
53. 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 Res. 2016;44:W58-63.
54. Nei M, Tajima F. DNA polymorphism detectable by restriction endonucleases. Genetics. 1981;97:145–63.
55. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536:419–24.
56. Browning BL, Browning SR. Detecting identity by descent and estimating genotype error rates in sequence data. Am J Hum Genet. 2013;93:840–51.
57. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet. 2016;48:811–6.
58. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31:2824–7.
59. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–8.
60. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:1–14.
61. Hothorn T, Zeileis A. partykit: A modular toolkit for recursive partytioning in R. J Mach Learn Res. 2015;16:3905–9.

Publisher's Note