The Impact of Recent Demography on Functional Genetic Variation in North African Human Groups

Marcel Lucas-Sánchez,[1] Amine Abdeli,[2] Asmahan Bekada,[3] Francesc Calafell,[1] Traki Benhassine,[2] and David Comas [1,*]

[1]Departament de Medicina i Ciències de la Vida, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain

[2]Faculté des Sciences Biologiques, Laboratoire de Biologie Cellulaire et Moléculaire, Université des Sciences et de la Technologie Houari Boumediene, Alger, Algeria

[3]Département de Biotechnologie, Faculté des Sciences de la Nature et de la Vie, Université Oran 1 (Ahmad Ben Bella), Oran, Algeria

***Corresponding author:** E-mail: david.comas@upf.edu.

Associate editor: Maria C. Ávila-Arcos

Abstract

The strategic location of North Africa has made the region the core of a wide range of human demographic events, including migrations, bottlenecks, and admixture processes. This has led to a complex and heterogeneous genetic and cultural landscape, which remains poorly studied compared to other world regions. Whole-exome sequencing is particularly relevant to determine the effects of these demographic events on current-day North Africans' genomes, since it allows to focus on those parts of the genome that are more likely to have direct biomedical consequences. Whole-exome sequencing can also be used to assess the effect of recent demography in functional genetic variation and the efficacy of natural selection, a long-lasting debate. In the present work, we use newly generated whole-exome sequencing and genome-wide array genotypes to investigate the effect of demography in functional variation in 7 North African populations, considering both cultural and demographic differences and with a special focus on Amazigh (plur. Imazighen) groups. We detect genetic differences among populations related to their degree of isolation and the presence of bottlenecks in their recent history. We find differences in the functional part of the genome that suggest a relaxation of purifying selection in the more isolated groups, allowing for an increase of putatively damaging variation. Our results also show a shift in mutational load coinciding with major demographic events in the region and reveal differences within and between cultural and geographic groups.

Key words: **North Africa, functional variation, whole exomes, human population diversity.**

Introduction

Throughout human history, North Africa has played a central role in many demographic movements because of its pivotal position in the intersection of 3 continents. Apart from cultural influences, these movements have brought to North Africa genetic inputs from all surrounding regions (the Middle East, Mediterranean Europe, and the rest of the African continent), which are present in the genomes of current North Africans along with an autochthonous genetic component (Henn et al. 2012; Arauna et al. 2017; Serra-Vidal et al. 2019; Lucas-Sánchez, Font-Porterias, et al. 2021; Lucas-Sánchez, Serradell, and Comas 2021). The proportions of these components vary from region to region, and even from population to population within regions, reflecting the complex and heterogeneous demographic history of North Africa (Arauna et al. 2017; Anagnostou et al. 2020; Lucas-Sánchez, Serradell, and Comas 2021). Culturally, North Africa also exhibits a complex mosaic of customs and traditions with 2 main cultural groups: Arabs and Imazighen (sing. Amazigh) (Camps 1996, 1998; Ghaki 2003; Pellat et al. 2012; Anagnostou et al. 2020). Imazighen are considered to descend from the Paleolithic inhabitants of North Africa (Camps 1995, 1998; Newman 1995; Fadhlaoui-Zid, Khodjet-el-khil, et al. 2011; Serra-Vidal et al. 2019; Fregel et al. 2018), who were later influenced at different degrees by Neolithic migrations to the region coming from Europe through trans-Gibraltar movements and from the Levant (Fregel et al. 2018; Simões et al. 2023). During historical times, Imazighen were deeply impacted at demographic and cultural levels by the Arab conquest of North Africa in the seventh century, and more importantly, by the large and relatively long-standing Bedouin immigration that followed the conquest in the eleventh century (Hiernaux 1975; McEvedy 1995; Newman 1995; Camps 1998; Anagnostou et al. 2020). Although a great proportion

Received: May 22, 2023. **Revised:** November 22, 2023. **Accepted:** December 19, 2023

of North Africans adopted the new culture, admixing with the arriving Arabs and beginning to identify themselves as such, some remained in or receded to isolated areas where the Amazigh culture, language, and identity are still maintained (Camps 1996, 1998; Masonen 1997; Idaghdour et al. 2010; Fadhlaoui-Zid, Khodjet-el-khil, et al. 2011). Within the Amazigh population, there is a rich diversity in both cultural and demographic terms, with a large range of different Tamazight languages (a branch in the large Afroasiatic family of languages spoken by Imazighen) and different histories involving isolation, admixture, nomadism, or sedentarism, and although they present some common elements, the different groups can be considered as different populations.

The recent demographic history of some Amazigh groups has been characterized by different degrees of genetic isolation and bottleneck events (Bosch et al. 1997; Camps 1998; Cherni et al. 2005; Fadhlaoui-Zid, Khodjet-el-khil, et al. 2011; Fadhlaoui-Zid, Rodríguez-Botigué, et al. 2011; Anagnostou et al. 2020; Lucas-Sánchez, Font-Porterias, et al. 2021; Lucas-Sánchez, Serradell, and Comas 2021). This might have had an impact on the genomes of current-day Imazighen and could allow us to trace back to historical events by applying genomic analyses, but like the vast majority of North African populations, the genomics of the Amazigh people are poorly studied. The impact of recent demography on human genomes is still a matter of debate. Studies in different populations, including a recent work in 2 North African populations (Lucas-Sánchez, Font-Porterias, et al. 2021), have shown that population groups with recent bottlenecks followed by small effective population sizes and genetic isolation present signs of relaxation of the purifying selection, probably driven by the increased effect of genetic drift and the lower genetic diversity. This results in an increase of the mutational load by the accumulation of deleterious mutations that would have been removed by purifying selection in larger and more genetically diverse populations (Lohmueller et al. 2008; Casals et al. 2013; Lim et al. 2014; Lohmueller 2014; Henn et al. 2016; Pedersen et al. 2017; Haber et al. 2019). However, other studies comparing Europeans and populations south of the Sahara show that they carry similar amounts of derived alleles, with no apparent effect of the out-of-Africa bottleneck in this regard, and suggest that genetic load is not affected by recent population size changes (Simons et al. 2014; Do et al. 2015). It is noteworthy that an important point of discrepancy is on how to measure the effect of selection, as no metric of genetic load or efficacy of selection has been universally accepted for human populations (Lohmueller 2014). Also important to note is the fact that studies showing a lack of effects of recent demography on genetic load are mainly based on comparing peoples of south of the Sahara to European groups, whose main bottleneck is relatively old (the out-of-Africa event), and who have experienced significant population growth since then (Boyko et al. 2008; Lohmueller et al. 2008; Fu et al. 2014; Lohmueller 2014; Simons et al. 2014; Do et al. 2015). Instead, studies on populations who have experienced a much more recent bottleneck, with large demographic effects and without important population growth, coincide in showing an accumulation of deleterious variants and homozygous derived genotypes, leading to an increased recessive load and a characteristic variant distribution with fewer rare alleles and more common ones, suggesting a relaxation of purifying selection (Casals et al. 2013; Lim et al. 2014; Pedersen et al. 2017; Lucas-Sánchez, Font-Porterias, et al. 2021). As mentioned above, some Amazigh groups carry a history of strong and recent bottlenecks, making them good candidates to study the genetic effects of recent demographic events, which can contribute to solve this debate.

In this context, in the present work, we have conducted a genomic characterization of 5 diverse Amazigh groups from different locations in the North African Maghreb, including the first in-depth genetic study of Chaoui (or Shawiya) populations from the Aurès, providing new whole-exome sequencing (WES) and genome-wide data and comparing them to geographically close Arab-speaking populations without a history of genetic isolation (more information on the newly sequenced populations can be found in Materials and Methods). Our analyses show a relevant impact of recent demography in the genomes of the studied Amazigh populations with potential biomedical effects, proportional to their level of isolation, and reveal historical patterns of mutational load with dates coinciding with major historical events in the region. Our results reflect the complex genetic landscape of North Africa, coherent with its demographic and cultural heterogeneity, and stress the need for more data and studies in a region usually neglected from genetic studies.

Results

Population Structure and Demographic Parameters

Principal component analysis (PCA) and ADMIXTURE analysis were performed to assess the population structure of the populations in the dataset, with a focus on the Amazigh samples (Fig. 1). In the first 2 PCA components, all North African samples cluster in the middle of a cline created in PC1 that goes from populations south of the Sahara to Europeans, with most samples being closer to the European reference population and with some of them overlapping with the Middle Eastern reference group. Nonetheless, some samples are placed much closer to the individuals from south of the Sahara, namely 5 Tunisian Arabs and 1 Mozabite. This might be a result of recent trans-Saharan gene flow (Lucas-Sánchez, Serradell, and Comas 2021; Lucas-Sánchez et al. 2023). Similar results were obtained when these samples were removed from subsequent analyses. A closer look at the North African cluster (plotting only North African individuals and removing outlier samples according to smartPCA criteria) reveals a clear differentiation of Tataouine Imazighen from the rest of North African samples in PC1, as well as internal heterogeneity in the Aurès region (Batna, Khenchela, and Oum El Bouaghi) in PC2, with most of

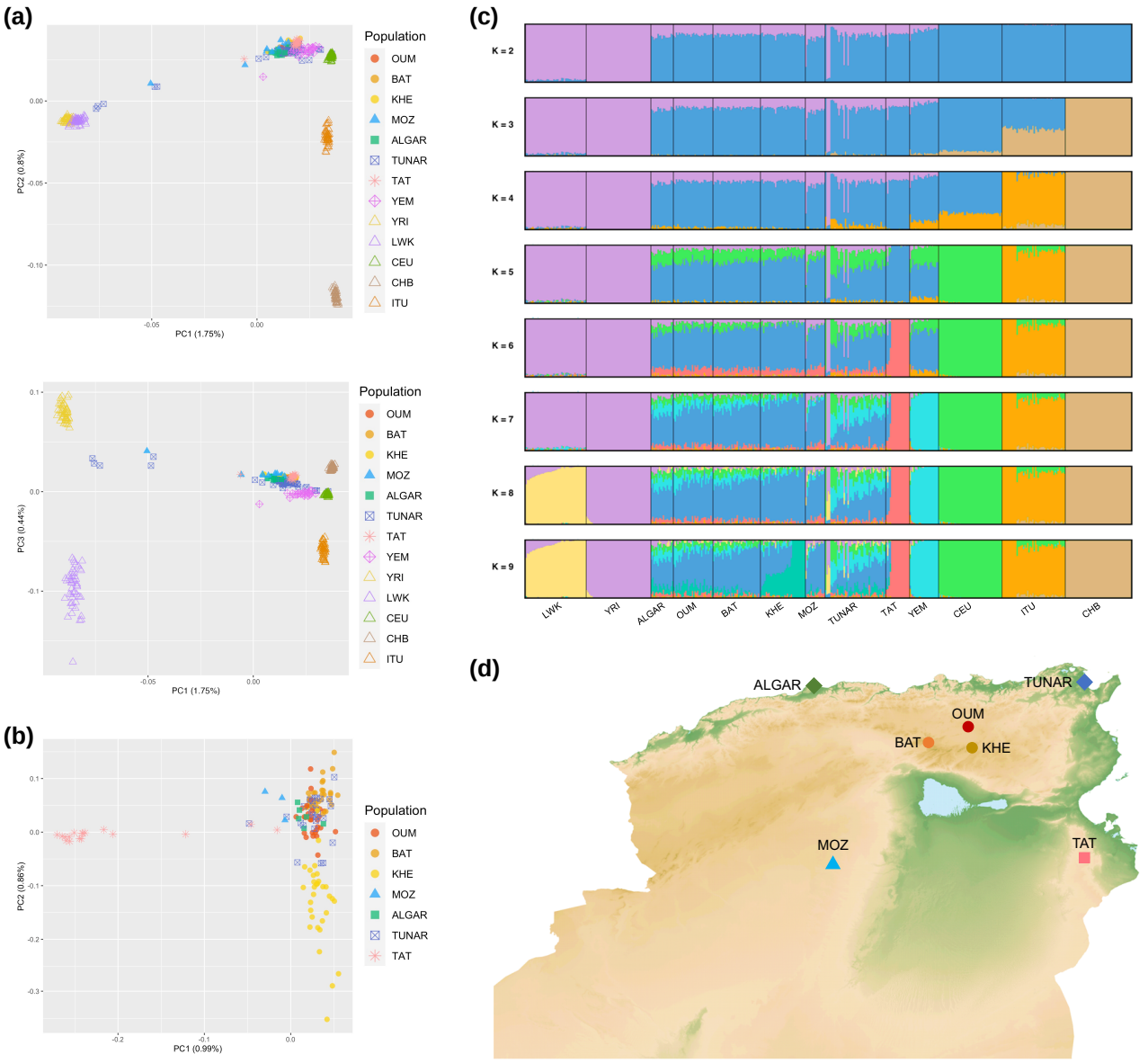


Fig. 1. PCA, ADMIXTURE analysis for $K = 2$ to 9, and sampled populations' locations for the whole-exome dataset. a) PCA of exome variants in 7 North African populations and a panel of relevant worldwide populations. b) PCA of the North African populations (excluding outlier samples). c) ADMIXTURE analysis ($K = 2$ to 9) of all populations present in a). d) Map of Tunisia and the northern half of Algeria with locations of the samples. Aurès populations are BAT, KHE, and OUM. Population abbreviations stand for Luhya (LWK), Yoruba (YRI), Algerian Arab (ALGAR), Oum El Bouaghi (OUM), Batna (BAT), Khenchela (KHE), Mozabite (MOZ), Tunisian Arab (TUNAR), Tataouine Imazighen (TAT), Yemeni (YEM), Utah residents with Northern and Western European ancestry (CEU), Indian Telugu in the UK (ITU), and Han Chinese in Beijing (CHB).

Khenchela samples separating from the rest in a clinal pattern (Fig. 1b). Mozabites appear close to but separated from the cluster formed by Arabs, Batna and Oum El Bouaghi.

The ADMIXTURE results (Fig. 1c) show a similar pattern to PCA, with $K = 5$ having the lowest cross-validation error (supplementary fig. S1, Supplementary Material online), but with $K = 2$ to 9 showing convergent results in more than half of the runs. Most of the ancestry of the Amazigh groups is composed of a North Africa-maximized component, with some Europe-maximized and, to a lesser extent, Western and Eastern Africa-maximized components also present. Tataouine Imazighen exhibit the most different proportions from all the groups and from $K = 6$ onward a Tataouine Imazighen-maximized component appears, comprising almost all of the ancestral proportions of Tataouine Imazighen and also appearing in the remaining North African groups. The Aurès groups present a very similar pattern through the first 8 *K*s, also similar to that of Algerian Arabs, and close to that of Tunisian Arabs and Mozabites, although these last 2 populations present some non-North African-maximized components at higher frequencies, with some samples having high proportions of the component maximized in populations south of the Sahara, concordant with PCA results. At $K = 9$, some Khenchela samples present a component at extremely

high proportions, being the only component in a third of the samples, mimicking the results of the North African-only PCA, and suggesting a local genetic heterogeneity within the Aurès region. A Middle Eastern-like component does not appear until $K = 7$, where it constitutes a relevant fraction of the ancestry of Tunisian Arabs, with lower presence in Algerian Arab and even lower in the Amazigh groups. Very similar results are found when using the merged (whole exome sequencing-single nucleotide polymorphisms [WES-SNP]) dataset (supplementary figs. S2 and S3, Supplementary Material online).

WES pairwise F_{ST} distances show matching results with PCA and ADMIXTURE, with Tataouine Imazighen being the most differentiated population and the Aurès populations being closer to both Arab populations than to Tataouine Imazighen and Mozabites. Mozabites themselves appear more similar to Aurès populations than to Arabs (supplementary fig. S4, Supplementary Material online). Within the Aurès, Khenchela appears again as the most differentiated, being Batna and Oum El Bouaghi the more similar ones. Khenchela and Batna, in fact, have lower F_{ST} distances with Oum El Bouaghi than between them. F_{ST} distances calculated with the 2 additional datasets show no differences to the WES analysis (data not shown).

The long-term N_e (the harmonic mean of N_e along the past generations explored), as studied with NeON (Mezzavilla 2015), shows relevant differences among Imazighen. Looking at general tendencies, Tataouine Imazighen present the lowest values followed by Mozabites, with which they partially overlap in the confidence intervals. Aurès populations present an intermediate N_e among the North African groups, with larger sizes than Imazighen from Tataouine or Mozabites, but smaller than the 2 nonhistorically isolated Arab groups, especially when compared to Tunisian Arabs (supplementary fig. S5a, Supplementary Material online). Within the Aurès, Khenchela exhibits a smaller size than the other 2 groups, and Batna a slightly smaller size than Oum El Bouaghi, although confidence intervals overlap. Similar results are found with the merged dataset (supplementary fig. S6, Supplementary Material online). The increase in a number of variants in this dataset clarifies some differences already suggested in the WES analysis (between Mozabites and the Aurès, between Khenchela, and both Oum El Bouaghi and Tunisian Arabs). In both datasets, Algerian Arabs exhibit N_e values close to and even a bit lower than Oum El Bouaghi and Batna. A possible explanation is the notably lower number of samples in this group. IBDNe (Browning and Browning 2015) results show similar patterns, with Aurès populations showing historical smaller sizes than Tunisian Arabs, but larger than Mozabites, especially in the last 40 generations (supplementary fig. S5b and c, Supplementary Material online).

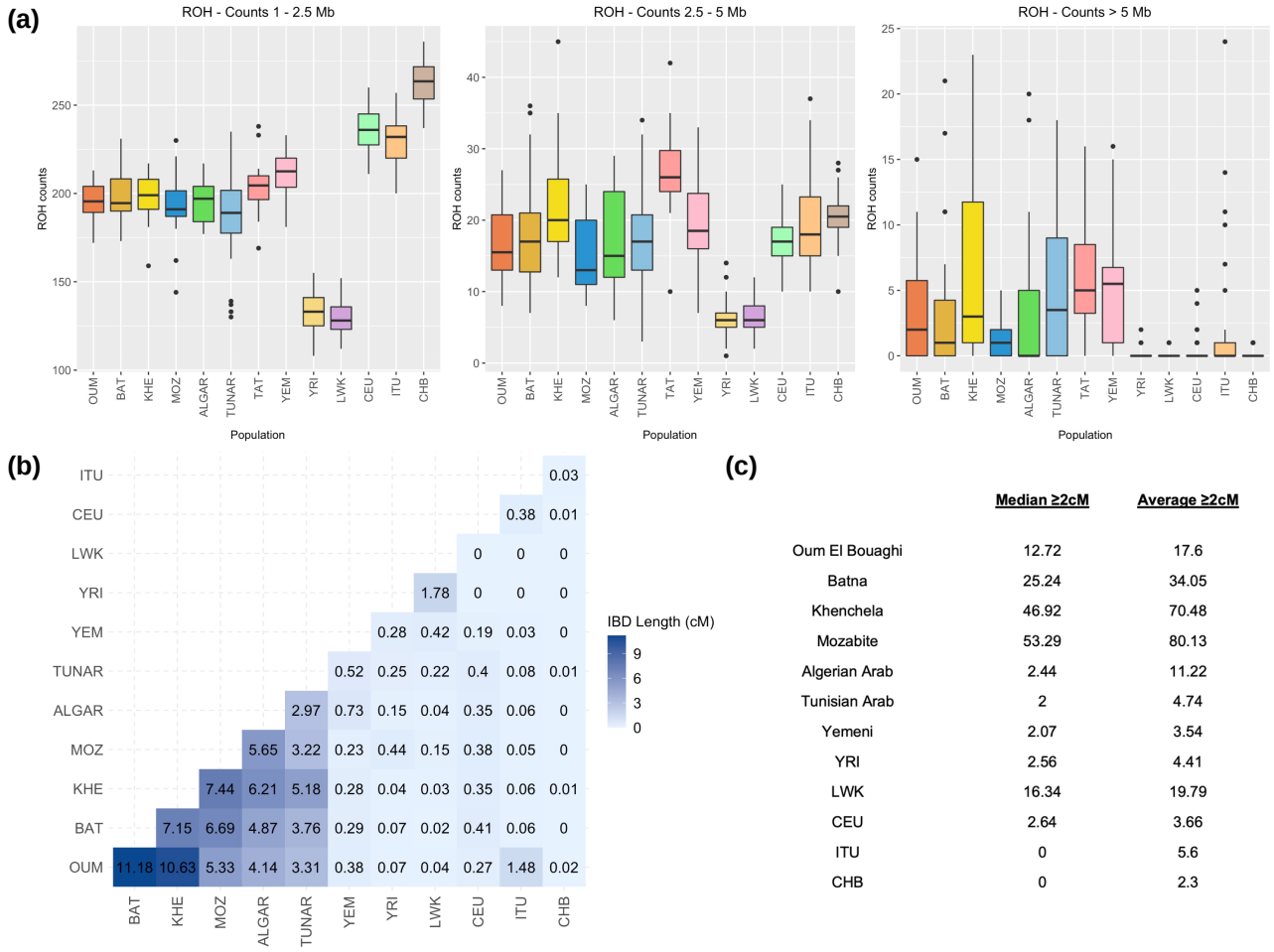
Genetic diversity indexes agree with N_e results, as those populations with lower N_e are also those with lower genetic diversity (supplementary fig. S7, Supplementary Material online) with the exception of Mozabites, in which the stronger trans-Saharan gene flow might be increasing genetic diversity despite having a small effective population size (see also supplementary Note S1, Supplementary Material online). Tataouine Imazighen have the lowest genetic diversity among the studied groups, and Aurès populations appear in an intermediate position between Tataouine Imazighen and the other North African populations, especially the Arab groups. Within the Aurès, Khenchela presents the lowest diversity values of the 3 and Oum El Bouaghi the largest.

ROHs and IBD Fragments Analyses

Further insight into inbreeding or consanguinity patterns was provided with the analysis of runs of homozygosity (ROHs) and identity-by-descent (IBD) segments. ROHs were divided by length categories as the interpretation depends on the length of those segments (longer runs are related to recent inbreeding and lower genetic diversity while shorter runs reflect past rather than recent small population sizes) (Ceballos et al. 2018). Similar to N_e and genetic diversity results, Tataouine Imazighen are the most divergent among Imazighen, exhibiting the largest ROH counts from all North African groups and, for runs 2 to 5 Mb long, the largest values in all populations. Khenchela is placed in a midway position between Tataouine Imazighen and the rest of the North African groups (Fig. 2a and supplementary tables S1 to S3, Supplementary Material online). Mozabites exhibit, for all categories, the lowest values in North Africa, only comparable to Batna for ROHs > 5 Mb. The analysis was repeated considering the total ROH length in each category with similar results (supplementary fig. S8 and tables S4 to S6, Supplementary Material online). The per-individual average ROH length also shows comparable results, with Tataouine Imazighen showing the largest mean values and Tunisian Arabs showing similar values to Khenchela in this analysis due to the presence of a few individuals with long ROHs (supplementary fig. S9 and table S7, Supplementary Material online). The merged dataset shows also similar results to the WES analyses (supplementary figs. S10 to S12 and tables S8 to S14, Supplementary Material online), with only an increase in Mozabite values probably driven by the exclusion from this dataset of some samples from the WES one (see "Materials and Methods" and supplementary Note S1, Supplementary Material online).

North African populations share more IBD segments between them than with other populations in the dataset, as expected, and among Imazighen, the higher IBD-sharing values are found between Aurès populations (it should be noticed that there is no genome-wide data available for the Tataouine Imazighen) (Fig. 2b). The 4 Algerian Amazigh populations (Aurès groups and Mozabites) share more IBD segments between them than with the 2 Arab groups and more than the 2 Arab groups between them.

The intrapopulation values are much larger than interpopulation ones, especially those found in the 4 Algerian



	Median ≥2cM	Average ≥2cM
Oum El Bouaghi	12.72	17.6
Batna	25.24	34.05
Khenchela	46.92	70.48
Mozabite	53.29	80.13
Algerian Arab	2.44	11.22
Tunisian Arab	2	4.74
Yemeni	2.07	3.54
YRI	2.56	4.41
LWK	16.34	19.79
CEU	2.64	3.66
ITU	0	5.6
CHB	0	2.3

Fig. 2. Runs of homozygosity and IBD fragments. a) Per-individual total count of runs of homozygosity in different length categories. Boxplots indicate the distribution of the per-individual total count of runs of homozygosity in each population. Points depict outlier individuals. Statistical significance can be found in supplementary tables S1 to S3, Supplementary Material online. Population names abbreviated as in Fig. 1. b) Average IBD segment length in cM between a random individual from the *x*-axis population and one from the *y*-axis population. Values are averaged over 1,000 independent comparisons. c) Median and average cumulative IBD segment length in cM between individual pairs in the same population.

Amazigh groups (Fig. 2c and supplementary fig. S13, Supplementary Material online). Mozabites and Khenchela show the highest values, suggesting higher levels of inbreeding and consistent with their lower effective population sizes (Fig. 2). Batna also presents high interpopulation IBD values. Oum El Bouaghi, although exhibiting the lowest values in the Aurès, still shows higher values than Arabs and most of the non-North African populations.

It is noteworthy that Mozabites are the population with the highest intrapopulation IBD values, but this tendency is not exactly matched by its ROH values. A possible explanation is that because of their low effective population size (supplementary fig. S5, Supplementary Material online), Mozabites have higher degrees of relatedness and inbreeding (hence the high intrapopulation IBD values), but their higher and relatively recent admixture with groups from south of the Sahara (Alport 1954; Harich et al. 2010) may have resulted in the introduction in the population of genomic stretches with less homozygous sites, while also splitting putatively previous runs.

Variant Distribution

The possible effects of recent demography in the genomes of Amazigh populations, and more specifically, its effects on the distribution of functional variants, were assessed with the analysis of the site frequency spectrum (SFS) (Fig. 3a and supplementary fig. 14a, Supplementary Material online) since it has been shown to be significantly affected by recent demographic history, genetic drift, and selection (Adams and Hudson 2004; Marth et al. 2004; Pedersen et al. 2017; Lucas-Sánchez, Font-Porterias, et al. 2021). Tataouine Imazighen, a population known to have been deeply affected by genetic isolation, clearly present the flattest SFS of all populations in the dataset, with a depletion of rare variants and a corresponding increase of variants in the more common categories, which has been interpreted as the effect of a potential relaxation of purifying selection (Lucas-Sánchez, Font-Porterias, et al. 2021). Regarding the Aurès groups, although all 3 groups present flatter SFS than the Arab and populations south of the Sahara, Oum El Bouaghi's SFS is close to that observed in Arab Algerians, while Khenchela presents the

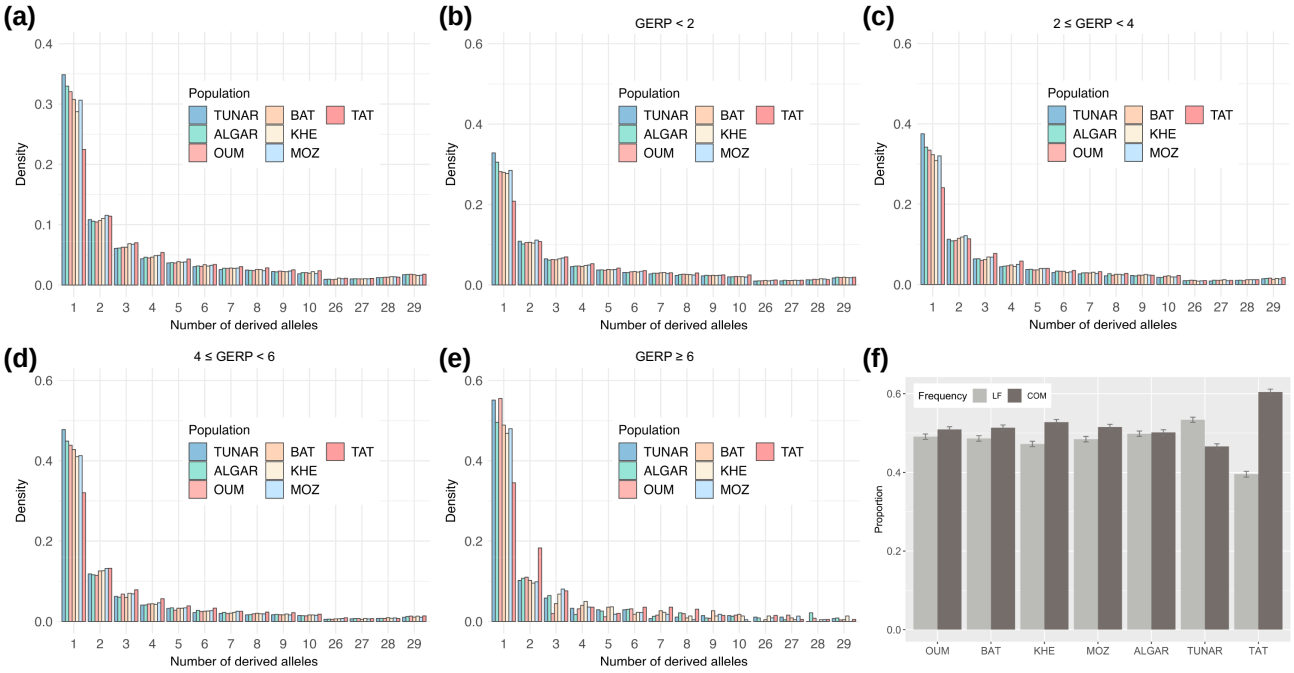


Fig. 3. SFSs of derived alleles and derived allele proportions. In all the analyses corresponding to these plots, populations were subsampled to 15 random individuals (for Mozabites, all the available individuals), and fixed sites were excluded. a) General SFS of derived alleles for all North African populations in the dataset. b to e) SFSs of derived alleles in different GERP RS score categories as indicated in the plot titles. f) Proportion of derived alleles predicted to be deleterious by GERP RS score (i.e. GERP RS scores > 2) classified by frequency category. Frequency-based categories include low frequency (LF), which accounts for singletons and doubletons, and common (COM), which accounts for tripletons and higher frequencies. Error bars represent the 95% confidence intervals. Population names in all plots abbreviated as in Fig. 1.

flattest SFSs in the Aurès. This points to a slightly lower effect of purifying selection in the Aurès populations than in the larger Arab groups, especially in Khenchela, probably because of an increase in genetic isolation and a decrease in effective population size, but not as strong as the one seen in Tataouine Imazighen. Mozabites show a similar pattern than Batna, especially for singletons, although in doubletons, they present the highest proportion in all the studied North African groups.

The patterns observed in the general SFS are retained when dividing the analysis by categories of Genomic Evolutionary Rate Profiling Rejected Substitution (GERP RS) score (Cooper et al. 2005; Davydov et al. 2010) (Fig. 3b to e and supplementary fig. 14b to e, Supplementary Material online), especially for the first 3 categories, which can be considered as neutral-, moderate-, and large-effect variants (Fig. 3b to d).

For a more concise visualization of these results, variants with GERP RS score ≥ 2, i.e. those considered to have a deleterious effect, were grouped by frequency in 2 categories: low-frequency variants (singletons and doubletons) and common variants (tripletons or higher), and these proportions were compared across North African groups (Fig. 3f). All Amazigh groups exhibit a higher proportion of their predicted deleterious variants in common frequencies than in low frequencies ($P < 0.005$), while Tunisian Arabs show the opposite pattern and Algerian Arabs show a statistically significant but very small difference in absolute values between both categories (low frequency = 0.49851, common = 0.50155, $P < 0.005$). Tunisian Imazighen show the steepest differences between proportions, followed by Khenchela, whose proportions are almost directly opposite to those of Tunisian Arabs.

Additional approaches to predict the deleteriousness of the variants (i.e. Combined Annotation Dependent Depletion [CADD] and PolyPhen scores) showed similar results (supplementary figs. S15 to S18, Supplementary Material online), although Batna, Oum El Bouaghi, and Mozabites show higher proportions of low-frequency deleterious sites predicted by CADD, and the lower number of variants predicted to be deleterious by PolyPhen compared to GERP and CADD causes an enrichment of rare variants in the PolyPhen analyses (as already observed in Lucas-Sánchez, Font-Porterias, et al. 2021), which still keep the relative interpopulation differences observed in GERP and CADD.

Mutational Load

Mutational load was estimated and compared between population pairs as an additional exploration of the burden and distribution of deleterious variations in North African Imazighen (Fig. 4). When assuming a fully additive model ($N_{alleles}$), the only significant comparisons (considering $P < 0.001$) are the higher values in all Amazigh populations compared to Yemeni and Europeans in the neutral category (GERP < 2) and lower values in Oum El Bouaghi than in Tunisian Arabs for those sites with $4 \leq$ GERP < 6. Under the assumption of a fully recessive model (N_{hom}),

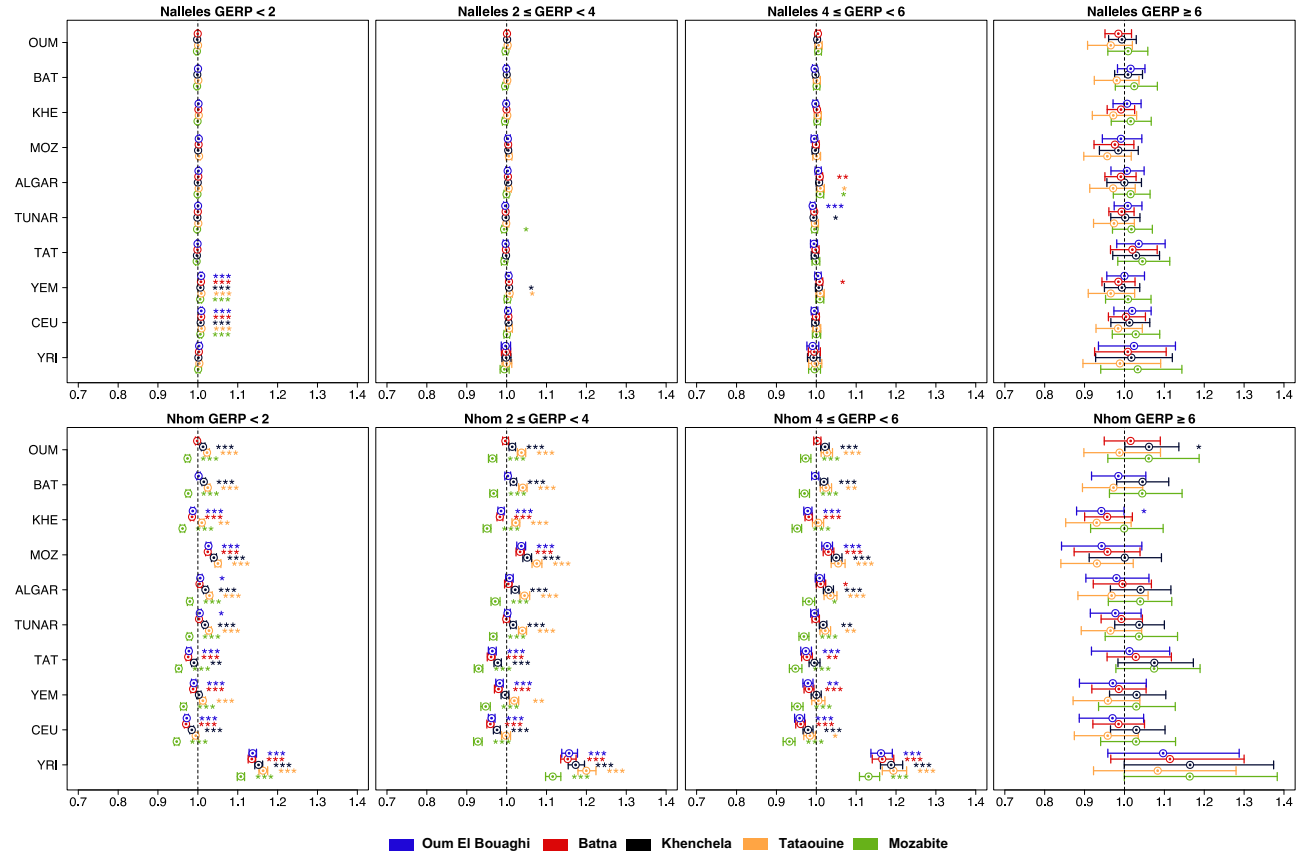


Fig. 4. Ratios of per-individual number of derived alleles ($N_{alleles}$, top row charts) and homozygous derived genotypes (N_{hom}, bottom row charts) between population pairs for variants in different GERP RS score categories as an approximation to mutational load comparisons. Each dot represents the ratio between an Amazigh population as depicted by the colors and the population written in the closest *y*-axis label. A ratio higher than 1 means more $N_{alleles}$ or N_{hom} in the population corresponding to the color of the ratio. A ratio lower than 1 means fewer $N_{alleles}$ or N_{hom} in the *y*-axis population. Plot titles indicate the range of GERP RS scores of the variants included. Error bars denote the 0.025 and 0.975 quantiles calculated after bootstrapping by site 1,000 times, sampling 1,000 random blocks of the exome data at each bootstrap iteration allowing resampling of the same block after dividing the data into 1,000 blocks. Statistical significance is shown in the following way: $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$. Significant values remain even correcting for multiple comparisons with a conservative Bonferroni correction (data not shown). Population names abbreviated as in Fig. 1.

Tataouine Imazighen present a significantly higher load than the rest of North Africans in the majority of comparisons, while Khenchela exhibits a significantly higher load than all North African groups different from Tataouine Imazighen. Batna and Oum El Bouaghi show similar values to the Arab groups, and Mozabites have significantly lower values than all other North Africans. The extreme category, which includes sites with GERP ≥ 6, has much fewer variants than the others (2,550 compared to 63,075 in the moderate category and 66,421 in the large category), causing the large confidence intervals observed in the ratios in this category. Selecting only GERP ≥ 2 variants labeled as missense in the Ensembl Variant Effect Predictor or using CADD or PolyPhen as deleteriousness prediction methods show similar results (supplementary figs. S19 to S21, Supplementary Material online).

In a similar line of analyses, we detected that in most populations, the ratio of deleterious (missense sites with GERP scores ≥ 2) to synonymous homozygous sites is higher inside ROH regions than outside, a tendency that is especially relevant in those sites with $4 \leq$ GERP scores < 6 and GERP scores ≥ 6 in the Aurès groups (supplementary figs. S22 to S24, Supplementary Material online). This enrichment in potentially deleterious homozygous variants inside ROH regions has already been revealed in previous studies (Szpiech et al. 2013; Ceballos et al. 2018; Font-Porterias et al. 2021; Lucas-Sánchez, Font-Porterias, et al. 2021) and indicates an increase in putatively damaging variation (and its possible biomedical consequences) in those populations with higher number and longer ROHs compared to those with less and shorter ROHs. This is the case when comparing Aurès populations to the Arab and non-North African populations in the present study. Other non-Amazigh groups in our dataset also present a significantly higher proportion of putatively deleterious variants inside ROH regions than outside, but it is not accompanied by a high number of long ROHs. Thus, the increase in damaging variation in these populations will not be as high as in other groups with a higher number of long ROHs, like some of studied the Amazigh populations. These results agree with those of the genetic load (Fig. 4).

One of the main goals of this work was to extend our analyses of the mutational load to study its historical patterns and temporal trajectory in the Amazigh groups through forward simulations. This requires a distribution of fitness effects (DFE) of new deleterious mutations, which in turn requires the choice of a demographic model for each population. In the absence of an established demographic model for the studied populations, several general models were tested for fitting with the observed data using ∂a∂i (Gutenkunst et al. 2009). The best-fitted demographic model for all the studied populations with suitable data for this analysis is a 3-epoch model (supplementary table S15, Supplementary Material online), with its corresponding demographic parameters (supplementary table S16, Supplementary Material online). On this basis, the DFE was estimated as a gamma distribution using the Fit∂a∂i software (Kim et al. 2017), resulting in a pair of scale and shape parameters for each population (supplementary table S17, Supplementary Material online). Both ∂a∂i and Fit∂a∂i estimations fit well with the observed data, i.e. the observed SFS and that expected from the neutral (∂a∂i) and selection (Fit∂a∂i) models are not significantly different (supplementary figs. S25 and S26, Supplementary Material online). DFE proportions are similar between populations, especially within the Aurès, with ∼30% neutral, ∼10% weakly deleterious, ∼12% to 14% moderately deleterious, and ∼45% to 50% strongly deleterious mutations (supplementary fig. S27, Supplementary Material online).

The final demographic model for each population entered in the forward simulations was estimated from ∂a∂i (DFE-related parameters) and IBDNe (effective population size trajectories) as explained in the Materials and Methods section, and 2 scenarios were tested: a fully recessive model (assuming all mutations have a recessive effect) and a fully additive model (assuming all mutations have an additive effect).

Under an additive model, mutational load for all populations keeps a relatively steady trajectory, close to Lg/Lanc = 1, with no major changes in the generations spanned by the simulations (Fig. 5). Contrastingly, mutational load under a recessive model increases significantly for all 4 Amazigh populations, starting around 35 generations ago in the Aurès (circa 10th to 11th century CE assuming 30 yr/generation; Tremblay and Vézina 2000) and around 40 generations ago in Mozabites (circa 9th century CE), being Khenchela the population with the strongest rise. Notice that these are not absolute mutational load values but values relative to an estimated ancestral mutational load. Algerian Arabs also present an increase of recessive load starting 15 generations ago, but lower than Batna, Mozabites, and Khenchela. It is noteworthy that in Oum El Bouaghi, Batna, and Mozabites, a slight decrease in additive load can be detected coinciding with the period of increase in the recessive model. Different numbers of generations for the burn-in phase were tested with similar results, and only a very small change was observed in Oum El Bouaghi, with the increase in recessive load happening later in the simulations and in a less pronounced way, which does not alter the interpretation of the overall results (data not shown).

Discussion

The genetic impact of recent demography and its possible effects on the efficacy of purifying selection are still a matter of debate. Some populations with low effective population sizes or affected by recent bottlenecks have been shown to accumulate slightly and moderately deleterious variation and homozygous derived genotypes, leading to an increase in mutational load and suggesting a decrease in the efficacy of purifying selection (Lohmueller et al. 2008; Casals et al. 2013; Lim et al. 2014; Henn et al. 2016; Pedersen et al. 2017; Lucas-Sánchez, Font-Porterias, et al. 2021). Isolation also seems to play an important role in the distribution of functional variation, as admixture (absent or low in isolated groups) has been shown to have an opposite effect on bottlenecks (Lopez et al. 2018; Font-Porterias et al. 2021). Additionally, no relevant effect of recent demography has been found in the amount of derived alleles in European populations when compared to sub-Saharan African groups (Simons et al. 2014; Do et al. 2015), implying that the population growth and contact with other populations happening in Europeans since the out-of-Africa bottleneck has counteracted the possible genetic impact of the bottleneck.

In this context, we present a study of the genetic impact of demography from a functional perspective in 5 different Amazigh populations, which have shown genetic evidence of bottlenecks, founder effects, and isolation events (Bosch et al. 1997; Cherni et al. 2005; Fadhlaoui-Zid, Khodjet-el-khil, et al. 2011; Fadhlaoui-Zid, Rodríguez-Botigué, et al. 2011; Henn et al. 2012; Anagnostou et al. 2020; Lucas-Sánchez, Serradell, and Comas 2021). We have compared them with each other and with surrounding non-Amazigh North African populations with different demographic histories. As the main differentiating point between these populations' demography is relatively recent, mainly in the times of the Arab expansion (seventh to eleventh centuries) (Ibn-Khaldoun 1968; Hiernaux 1975; McEvedy 1995; Newman 1995; Anagnostou et al. 2020; Lucas-Sánchez, Font-Porterias, et al. 2021), our results show the impact of recent demography on the genomes of the studied populations. North Africa is poorly represented in population genetic studies, and even less represented are Amazigh populations despite having a complex demographic history that makes them interesting candidates for these studies (Lucas-Sánchez, Serradell, and Comas 2021).

Our results, based on whole exome- and genome-wide data, show that Amazigh populations present genetic differences with populations that have not experienced recent bottleneck and isolation events. Besides, genetic differences can also be observed between different Amazigh groups in patterns that can be related to the predicted strength of a relatively recent bottleneck event and to the resulting population size and degree of genetic isolation. Imazighen from Tataouine show the strongest signs

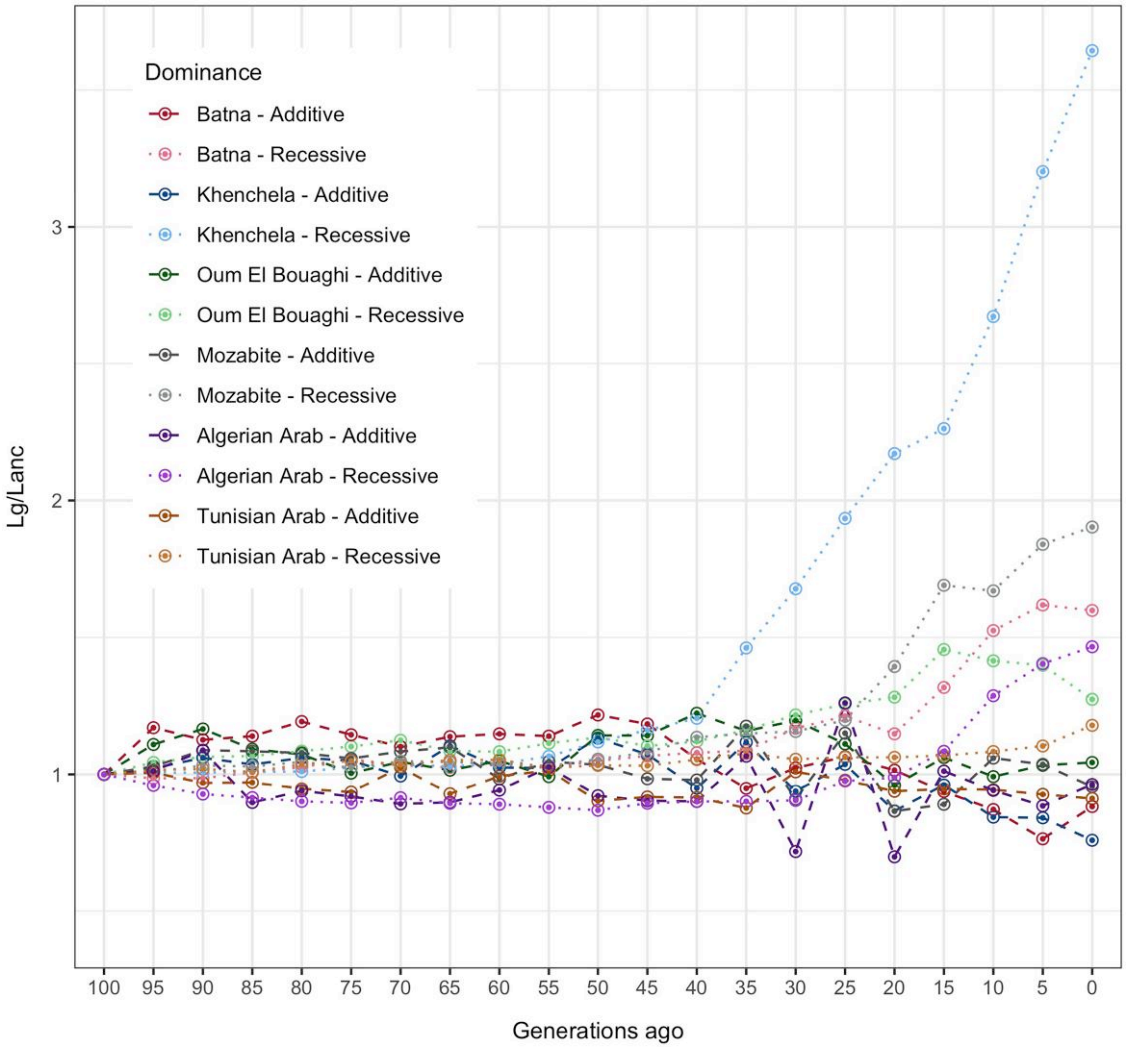


Fig. 5. Trajectory of mutational load of North African populations relative to the estimated ancestral load 100 generations ago (Lanc). The *y*-axis shows the relative mutational load (Lg/Lanc) calculated at each sampled generation (*x*-axis). The demography of each population is taken from IBDNe results and the genetic parameters are calculated from the DFE estimated with ∂a∂i.

of genetic isolation, drift, and inbreeding, with an increase in the proportion of variants found at high frequencies and a depletion of rare variants, a tendency that it is maintained when checking putatively damaging variants and leads to higher mutational load than any other population in the dataset (under a recessive model). These results reinforce the idea that recent demographic events in this population have led to a decrease in the efficacy of purifying selection, allowing for possibly deleterious variants to reach higher frequencies than they would have reached in a population with a different demographic history, which is in agreement with Lucas-Sánchez, Font-Porterias, et al. (2021). Tataouine Imazighen come from 2 very small villages in the desertic south of Tunisia, set in a region with cliffs and hills, which complicate accessibility and favor isolation. Although historical information and records are scarce, it is possible that these populations were affected by a strong bottleneck during the Arab expansion, receding to these remote locations to maintain an Amazigh identity as happened to different Amazigh groups (Camps 1995, 1998; Masonen 1997; Idaghdour et al. 2010; Fadhlaoui-Zid, Khodjet-el-khil, et al. 2011), which would explain the obtained results.

The Amazigh groups from the Aurès appear in an intermediate position in the analyses related to genetic isolation, drift, and inbreeding, presenting stronger signs of isolation than the 2 larger Arab groups (especially the Tunisian Arabs) but not as strong as those found in the Tataouine Imazighen. Results of effective population size and genetic diversity agree with the Aurès groups being relatively small populations with a stronger effect of genetic drift and inbreeding than other larger populations, but far from other very isolated Amazigh groups. Compared to the non-isolated North African groups in the dataset, Aurès populations exhibit more and longer ROHs, larger interpopulation IBD segments, and a shift of the SFS toward higher frequencies, a tendency they share—although with differences—with the other studied Imazighen. The Mozabites are a particular case among the Amazigh groups in the present

analysis. They present the second-lowest effective population size in the dataset and comparable SFS results to Batna, but they have a significantly lower number of ROHs and a significantly lower recessive load than the other studied Imazighen. A possible explanation is that Mozabites have experienced a strong bottleneck and genetic isolation comparable to other Amazigh groups, resulting in low effective population size values, high IBD sharing, and a relatively flat SFS, but recent contacts with peoples coming from south of the Sahara (Alport 1954; Harich et al. 2010) have introduced genetic segments that have break possible ROHs, increased genetic diversity values, and have in consequence lowered the recessive load values. The Mozabite samples included in this work present heterogeneity regarding the proportion of their ancestry related to populations south of the Sahara, which is expected in a recent admixture scenario. Nonetheless, its impact in some results is worth mentioning. The clearest case is in the estimation of the effective population size, in which separating the samples between their 2 sources (newly genotyped and Henn et al. 2016), which roughly coincides with their proportions of ancestry related to populations south of the Sahara, results in 2 lower N_e values than the merged Mozabite group, as this later group has higher diversity than the 2 separate ones, increasing N_e. Additional information and discussion in this regard can be found in supplementary Note S1, Supplementary Material online.

In all North African populations, variant distribution results show similar results as those discussed above when the analyses are focused on the variants with putatively damaging effects, indicating possible biomedical implications of the observed differences. This agrees with the analyses showing an excess of potentially deleterious homozygous variants inside ROH regions compared to the rest of the genome and the fact that long ROHs have been shown to be involved in predisposition to rare and common diseases (Ceballos et al. 2018). This is especially relevant in the Amazigh groups from the Aurès and Tataouine, in which a reduction of effective population size caused by a bottleneck followed by genetic isolation may have led to an increase in inbreeding and, as a consequence, of the number of ROHs and the consequent increase in potentially damaging homozygous variants.

Imazighen do not only show differences among different regions but also even between populations in the same region, as seen in the Aurès groups, which are geographically very close (in an area of ~3,000 km^2). Although in most analyses, and notably in the interpopulation IBD calculations, the Aurès populations present more similarities between them than with the remaining North African groups, they also present relevant internal differences. Khenchela appears as the group with stronger signs of isolation, genetic drift, and inbreeding, with higher number of ROHs, a flatter SFS, lower effective population size, and larger interpopulation IBD values, among other results. This differentiation is especially relevant in the results of mutational load, which show that, under a recessive model, Khenchela present a higher load than the 2 other Aurès groups for the first 3 GERP RS score categories. Khenchela also presents a higher load than both Arab groups and Mozabites, while Batna and Oum El Bouaghi's load is only significantly higher than that of Mozabites. The remaining analyses also show concordant outcomes, with Khenchela being the population within the Aurès with more divergent results, in both variant distribution and population structure analyses. Interestingly, Khenchela is not the population with the lowest census size as one might expect given these results, which is in fact Oum El Bouaghi (Office National de Statistique 2009). Other factors may be driving this differentiation in the Aurès, being geography and accessibility a possible explanation given the fact that Khenchela has the highest elevation of all 3 Aurès locations and is located in a more mountainous region than Oum El Bouaghi. Batna, although being located in the center of the Aurès mountains, has historically acted as the capital of the Aurès and has been better connected than other locations in the mountains (Kimble 1941).

Previous studies in North Africa have revealed a significant presence of interfamily marriages among its inhabitants, in different degrees depending on the location, a practice that can increase homozygosity in the following generations (Anwar et al. 2014, Bachir and Aouar 2019). Studies of this type have been performed in different locations from Algeria and Tunisia, but no data are available for the populations of interest of the present work located in the Aurès and Tataouine regions. However, although this is an aspect to consider when studying North African populations, we do not expect a significant cultural difference between Arabs and Imazighen regarding interfamily marriages, and hence, we do not believe it affects their relative results.

We present in this work the first study of the temporal trajectory of mutational load performed in North African populations. We observe a relevant change in mutational load under a recessive model in the Amazigh groups included in the analysis at comparable time points. In the Aurès, the change started at around 35 generations ago with an increase of load especially large in Khenchela and a coinciding decrease in additive load more noticeable in Oum El Bouaghi and Batna. In Mozabites, we observe a slight increase starting at around 40 generations ago that changes to a steep rise in the last 20 to 25 generations (the first generations of which coincide with a decrease in additive load). This is not observed in Tunisian Arabs, and only in part in Algerian Arabs, which present an increase in recessive load at a much more recent time. The changing point in load in the Aurès groups (less clear but also possible in Mozabites) roughly coincides with the time of the large Bedouin migration to North Africa, which happened in the 11th century CE (Hiernaux 1975; McEvedy 1995; Newman 1995; Camps 1998), which assuming 30 yr/generation is around 34 generations ago (Tremblay and Vézina 2000). This was the largest demographic movement in the context of the Arab expansion, and one of its consequences was the receding of some

Amazigh populations to remote and mountainous regions like the Aurès, reducing their effective population sizes and increasing genetic isolation (Ibn-Khaldoun 1968; Camps 1996, 1998; Fadhlaoui-Zid et al. 2004; Idaghdour et al. 2010; Fadhlaoui-Zid, Rodríguez-Botigué, et al. 2011; Anagnostou et al. 2020). This could have led to an increase in mutational load, mostly by the accumulation of potentially damaging variants in homozygous form (thus, recessive load) as a consequence of the reduced genetic diversity and increased inbreeding. It is noteworthy that previous studies have shown a counteracting effect of gene flow on mutation load in similar simulations as the ones here performed (Font-Porterias et al. 2021). We did not consider gene flow in our simulations as we would need a demographic model of the studied populations, currently unavailable, and although gene flow could in fact have reduced the increase in genetic load we detect, the increase would likely not be completely compensated because of the situation of relative genetic isolation of these populations and the fact that a major impact of gene flow post-Arab expansion is not expected until very recent times (Ibn-Khaldoun 1968; Camps 1996, 1998; Fadhlaoui-Zid et al. 2004; Idaghdour et al. 2010; Fadhlaoui-Zid, Rodríguez-Botigué, et al. 2011). Mutational load trajectory results are concordant with those observed in our present-day mutation load analyses, where we see differences in mutation load under a recessive model but not under an additive model.

Interestingly, while Mozabites present an increase in mutational load in the last 40 generations, their present-day load is still lower than the populations in the Aurès under both models of dominance, suggesting that although they may have experienced a recent reduction in effective population size and an increase in genetic drift, other recent factors such as possible admixture with peoples from south of the Sahara (Alport 1954; Harich et al. 2010) (a consequence of the historical presence of slave-trade routes in Mozabite territory and their closeness to the desert) might have acted as a compensation leading to a lower present-day load than other Amazigh populations with lower admixture. This would be concordant with the suggested counteracting effect of admixture on genetic load (Font-Porterias et al. 2021).

Our results have implications in the debate of the effect of recent demography in the efficacy of purifying selection, as we observe relevant differences in the distribution of functional variants in population groups with different demographic histories and in concordance with their degree of isolation and demographic impact of recent historical events. Populations with smaller effective population sizes and a potentially stronger bottleneck are predicted to have lower genetic diversity (Wright 1937; Muller 1950; Morton et al. 1956; Kimura et al. 1963) and thus less variation for purifying selection to act upon. This, coupled with the increase of genetic drift, which acts oppositely to purifying selection (Haldane 1927, 1937; Wright 1937; Kimura et al. 1963), results in a decrease of the efficacy of the latter in purging damaging variation and the corresponding increase of such variants observed in the present study. Our analyses align with these hypotheses, suggesting that isolation in the studied Amazigh groups after a relatively recent bottleneck has lowered the genetic diversity and resulted in an increase of high-frequency variants both in general and for potentially damaging variation, pointing to a decrease in the efficacy of purifying selection in removing these variants and an increase in the effect of genetic drift. Moreover, results of the temporal trajectory of mutational load point to an important effect in the genomes of these populations as a possible consequence of a major demographic event in the region.

Although this study includes a panel of 7 North African groups, North Africa is a culturally and demographically complex and highly heterogeneous region, and even within a relatively small region like the Aurès, one can find groups with relevant cultural and genetic differences (as illustrated here with the population of Khenchela), and to this date, genetic data in North Africa are still scarce. We believe that, while recent efforts are increasing the availability of genetic data in the region (Arauna et al. 2017; Serra-Vidal et al. 2019; Anagnostou et al. 2020; Lucas-Sánchez, Font-Porterias, et al. 2021; Lucas-Sánchez et al. 2023), more data on these and different populations, covering both the geographical and cultural diversity of North Africa, should be gathered in order to fully understand its genetic landscape and provide substantial biomedical benefits to the population. The results of the present work also show the high heterogeneity among Amazigh groups and stress the need to consider them as separate populations in genomic and biomedical studies.

Materials and Methods

Ethics Declaration

Written informed consent was obtained from all the volunteers, and the present project has the corresponding IRB approvals (*Comitè d'Ètica d'Investigació-Parc de Salut Mar* 2019/8900/I, Barcelona, Spain; and *Université des Sciences et de la Technologie Houari Boumediene*, Algiers, Algeria). All the methods were carried out in accordance with relevant guidelines and regulations.

Sample Collection

Samples from 3 Chaoui or Shawiya Amazigh populations in the Aurès region (Algeria) were newly sequenced in the present study and collected in the towns of Batna ($n = 38$), Khenchela ($n = 37$), and Oum El Bouaghi ($n = 35$). These 3 locations belong to the wilayas (provinces) of the same name, all located in the natural region of the Aurès, in the northeastern part of Algeria, characterized by its mountainous terrain and historically relatively difficult accessibility compared to other parts of the country (Kimble 1941). Batna acts as the capital of the Aurès. Their census sizes in latest official census are 290,645 inhabitants in Batna, 108,580 in Khenchela, and 80,359 in

Oum El Bouaghi (Office National de Statistique 2009). Additionally, samples from 2 other Algerian populations were also newly sequenced: Mozabite Amazigh individuals from Ghardaïa ($n = 8$), in the wilaya of the same name, situated in the M'zab valley in central Algeria, and with a census size of 93,423 inhabitants (Office National de Statistique 2009), and Arab individuals from the city of Algiers ($n = 32$), the capital of the country, situated in the Mediterranean coast and with a census size of 2,988,145 inhabitants in the latest census although it may have grown since then (Office National de Statistique 2009). Participants were healthy volunteers with appropriate informed consent and volunteers were informed about the aim of the project, which is focused on the genetic characterization of the Algerian population, and members of the research team were answering the questions of the volunteers during and after the sampling process. Aurès individuals were Chaouïa speakers (or Shawiya), a language in the Tamazight family, who live and were born in the assigned locality and whose parents and 4 grandparents were born in the region. Mozabite individuals were Mozabite speakers, another language in the Tamazight family, while Arab individuals were Arab speakers. Details of the project and preliminary data on the population genetic analyses of the present samples were publicly presented to the general audience during the PhD thesis defenses of members of the team (Bekada 2015; Abdeli 2021). Samples were collected for nonrelated individuals in these populations, and WES was performed with the Agilent SureSelect Human All Exon V6 capture kit.

Data Preparation and Quality Controls

FASTQ files containing raw sequencing data were first trimmed to remove adaptor sequences with Trimmomatic (Bolger et al. 2014). Before and after trimming, read quality was assessed with the fastqc online tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Variant calls from FASTQ files were obtained following the Genome Analysis Toolkit (GATK) Best Practices recommendations (Van der Auwera et al. 2013). Read pairs were mapped to the GRCh37 human reference genome with the Burrows–Wheeler Aligner (BWA) version 0.7.15 (Li and Durbin 2009). Mapped reads were merged, and PCR duplicates were removed using the MarkDuplicates tool from Picard 2.18 (https://broadinstitute.github.io/picard/). GATK 3.7 (McKenna et al. 2010) was used to perform quality control on mapped files (BAM) regarding coverage and mapping. This was performed before and after duplicate removal. The same software was used for indel realignment and base quality score recalibration with the following tools (in order): RealignerTargetCreator, IndelRealigner, BaseRecalibrator, and PrintReads. The HaplotypeCaller tool was then used for SNP and indel discovery. Additional samples were included to the dataset at this point, both from North Africa and from different worldwide regions, to serve as reference for the analyses. From North Africa, 18 Tunisian Amazigh samples from the Tataouine region and 46 Tunisian Arab samples from Tunis previously published were included (Lucas-Sánchez, Font-Porterias, et al. 2021), as well as 8 additional Mozabite samples (Henn et al. 2016). The available files for these Mozabite samples were mapped to the hg19 human reference genome, which presents some slight differences from GRCh37 mainly in some labels, so a custom-modified version of the hg19 reference was used to match the GRCh37 reference for the processing of Mozabite samples. Published Yemeni samples were also included (Haber et al. 2019). These were 47 whole-genome samples mapped to the GRCh38 human reference genome, so after obtaining the corresponding filtered and processed BAM files, lift over to the GRCh37 reference was performed using CrossMap (Zhao et al. 2014). Finally, a panel of 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) populations were also included, randomly selecting 50 individuals from each of the following: Yoruba in Ibadan, Nigeria, (YRI) from Western Central Africa; Luhya in Webuye, Kenya, (LWK) from Eastern Central Africa; Utah residents (CEPH) with Northern and Western European ancestry (CEU); Han Chinese in Beijing (CHB); and Indian Telugu in the UK (ITU). SNP and indel calling was performed in these samples in the same way as in the newly sequenced. Then, individual variant files for all the stated samples were combined using the GenotypeGVCF tool from GATK 3.7, resulting in a single variant call format (VCF) file for the complete dataset. In this joint variant calling, Yemeni whole-genome samples retained only those sites intersecting with the exome variants. VariantRecalibrator and ApplyRecalibration GATK tools were used to recalibrate SNPs and indels. Variant quality control was performed based on the guidelines in Lopez et al. (2018) and using VCFtools 0.1.14 (Danecek et al. 2011). Variants and samples were excluded following the same criteria as in Lucas-Sánchez, Font-Porterias, et al. (2021), removing a total of 28 samples. Before and after applying these filters, VCF quality controls were performed.

Individual relatedness was assessed to complement the relatedness-avoiding sampling strategy in the same manner as in Lucas-Sánchez, Font-Porterias, et al. (2021). Quality controls were performed after all data preparation steps. As an internal control, 4 samples were independently sequenced and filtered twice, giving a mean correspondence of 99.99% between duplicate samples, and so corroborating the robustness of the data preparation process.

The final dataset contained 366,407 SNPs with a mean coverage of 55.97X, and a mean missingness of 0.45%. Ti/Tv ratio was 2.78, which is expected in high-quality variant datasets (Carson et al. 2014). The final sample count per population was as follows: Batna $n = 36$, Khenchela $n = 34$, Oum El Bouaghi $n = 30$, Algerian Arabs $n = 17$, new Mozabite samples $n = 7$, previous Mozabite samples $n = 8$, Tataouine Imazighen $n = 18$, Tunisian Arabs $n = 46$, Yemeni $n = 22$, YRI $n = 49$, LWK $n = 46$, CEU $n = 48$, ITU $n = 48$, and CHB $n = 50$, for a total of 459 individuals (supplementary tables S18 and S19, Supplementary Material online).

Additional Datasets

Two additional datasets were used in the present study. First, a second WES dataset containing only North African populations was created, allowing us to filter out all missing sites while maintaining a significant number of sites (301,804 or 287,404 before and after ancestral state determination, respectively) and increasing coverage to 59.8X. This was used to replicate some analyses with and without missing sites, accounting for the possible differences that missing sites could create.

Another dataset was created by merging WES data and genome-wide SNP array data for populations in which both data types were available. The newly exome-sequenced samples were also genotyped with the Affymetrix Axiom Genome-Wide Human Origins 1 array. Genotype calling was performed using the Axiom Analysis Suite 4.0.3 software with default parameters and thresholds. PLINK 1.9b (Purcell and Chang; Chang et al. 2015) was used to keep only autosomes and biallelic sites, remove duplicate SNPs and indels, and filter out variants with >5% missingness. All variants passed the Hardy–Weinberg exact test ($P < 10^{-5}$). Related individuals were filtered out in the same way as in the WES dataset, and all samples passed the filter of <5% individual missingness. This dataset was merged with genome-wide array data corresponding to the individuals in the WES dataset from the Tunisian Arab population (Lucas-Sánchez et al. 2023) and with the corresponding genome-wide sites extracted from the whole-genome Yemeni and 1000 Genomes Project samples that were used in the WES dataset. Tunisian Imazighen were not included in this dataset because genome-wide array data are not available for these samples. The merging of genome-wide datasets was performed with VCFtools, keeping only those sites shared between all populations, resulting in a dataset of 475,919 variants. Finally, the genome-wide and the WES datasets were merged, concatenating and sorting both VCF files and keeping only those individuals present in both datasets. A final set of quality controls was applied to filter out sites with >5% of missingness and fixed sites, resulting in a final variant number of 806,255. Additional filters to filter out samples with individual missingness <5% and SNPs failing the Hardy–Weinberg test at $P < 10^{-5}$ did not remove any samples nor sites. The final sample count in each population was as follows: Batna $n = 28$, Khenchela $n = 28$, Oum El Bouaghi $n = 30$, Algerian Arab $n = 16$, new Mozabite samples $n = 7$, Tunisian Arab $n = 41$, Yemeni $n = 22$, YRI $n = 48$, LWK $n = 46$, CEU $n = 48$, ITU $n = 48$, and CHB $n = 50$, for a total of 412 individuals (supplementary table S18, Supplementary Material online). This dataset was used in analyses that do not consider the functionality of variants and those that need nonexonic variants or a higher number of variants than those in the whole-exome dataset. It was also used to repeat analyses performed with the whole-exome dataset. All analyses using IBD segments, as well as the trajectory of mutational load and its related analyses, were performed using only the merged dataset (supplementary table S19, Supplementary Material online).

Population Structure Analyses

For PCA and ADMIXTURE analysis, data were first pruned for linkage disequilibrium using PLINK 2.0 (Purcell and Chang; Chang et al. 2015) with the same parameters as in the relatedness estimation. This kept 298,380, 214,306, and 526,267 sites in the complete WES dataset, the WES dataset with only North Africans, and the merged WES-SNP dataset, respectively. The SmartPCA tool from the EIGENSTRAT stratification correction method implemented in the EIGENSOFT software package version 6.0.1 (Patterson et al. 2006) was used to perform PCA (Fig. 1a and b). Some samples were detected as outliers by the algorithm and removed from the analysis. ADMIXTURE 1.3 (Alexander et al. 2009) in unsupervised mode was applied to explore ancestry patterns in our dataset, with a number of ancestral clusters ranging from $K = 2$ to 10 and performing 50 independent runs with different random seed values for each K (Fig. 1c and supplementary fig. S2, Supplementary Material online). Data were filtered to remove variants with frequencies lower than 0.05 (Minimum allele frequency = 0.05). Cross-validation errors were assessed at each run and mean values were calculated to determine the minimum error range (supplementary figs. S1 and S3, Supplementary Material online). Common modes among the different runs for each K were identified using pong in greedy mode (Behr et al. 2016), which also allows for visualization and plotting of the results. Intermediate files for these analyses were obtained using PLINK 2.0 (Purcell and Chang; Chang et al. 2015) and VCFtools 0.1.14 (Danecek et al. 2011).

F_{ST} values between population pairs were computed using VCFtools 0.1.14 (Danecek et al. 2011) in all 3 datasets (supplementary fig. S4, Supplementary Material online). Comparisons were made using all individuals in each population.

Ancestral State Allele Determination

The ancestral and derived state of alleles at each site was determined in the 2 WES datasets following the approach described in Lopez et al. (2018), using the 6-EPO multialignment from Ensembl Compara version 59. Sites with unknown ancestral allele were removed from all WES analyses apart from those described in the population structure analyses section, keeping 351,994 sites in the complete WES dataset and 287,404 sites in the WES dataset with only North Africans.

Effective Population Size Estimations

Effective population sizes (N_e) over time were estimated with 2 approaches: NeON (Mezzavilla 2015) in the complete WES and in the merged datasets (supplementary figs. S5a and S6, Supplementary Material online, respectively) and IBDNe (Browning and Browning 2015) only in

the merged dataset (supplementary fig. S5b and c, Supplementary Material online). NeON is an R-package that bases N_e calculations in linkage disequilibrium patterns and has previously been used in WES data in Nutile et al. (2019) and Lucas-Sánchez, Font-Porterias, et al. (2021). It was used in the present study to calculate long-term effective population size, which is the harmonic mean of N_e along the generations in the past. Each chromosome was used as a replicate to calculate the mean (percentile 50th), and the confidence intervals (percentiles 5th and 95th), as provided by the Ne_CI function in the NeON package.

IBDNe is based on IBD fragments and required phased haplotypes. Thus, the merged dataset was first phased using Beagle 5.3 (Browning et al. 2021), and IBD segments were detected using the haplotype-based IBD detection method refine IBD (Browning and Browning 2013) with default parameters and with a 2-cM IBD segment-length threshold. Possible gaps between IBD segments resulting from genotype and haplotype phase errors were filled using merge IBD (Browning and Browning 2013). Gaps were filled if they were <0.6 cM long and there was a maximum of 1 pair of genotypes inconsistent with IBD in the gap. IBD segments at least 2 cM long were then inputted to IBDNe to calculate the N_e trajectories. To calculate the confidence intervals, we used the default 80 bootstrap replicates, which are performed by resampling chromosomes with replacement.

IBD Sharing

IBD segments were calculated as explained in the Effective Population Size Estimations section. Population pairwise comparisons were computed by randomly selecting an individual for each population and adding all IBD segments between both individuals (Fig. 2b). This was performed 1,000 times for each population pair (including within-population individual pairs), and average values were calculated, to account for variability. Within-population median and average IBD sharing were calculated by averaging across all IBD segments in the corresponding population (Fig. 2c and supplementary fig. S13, Supplementary Material online).

Genetic Diversity

Diversity indexes were calculated for all populations in the 2 WES datasets based on the SFS for the synonymous sites, as in Lopez et al. (2018) and Lucas-Sánchez, Font-Porterias, et al. (2021) (supplementary fig. S7, Supplementary Material online). Four different statistics were computed: pairwise nucleotide diversity ($\theta\pi$), Watterson's estimator (θw) and Tajima's D based on Kousathanas et al. (2011), and the nucleotide diversity for variable sites (πvar) based on Pedersen et al. (2017). Bootstraps by site were performed to calculate confidence intervals and statistical significance, with 1,000 replicates dividing each time the exome into 1,000 blocks and randomly keeping 1,000 of them allowing resampling. This approach considers the possible variance introduced by demographic processes (Gravel 2016; Simons and Sella 2016). Statistical significance was calculated with pairwise t-tests, and confidence intervals were set as the 0.025 and 0.975 quantiles of the bootstrap distribution.

As this is based on SFS, in the complete WES dataset, this was performed (i) separating the 2 Mozabite sample groups (the newly sequenced and those from Henn et al. 2016), thus taking 7 random individuals per each population, and (ii) grouping them together, thus taking 15 random individuals per population (see supplementary Note S1, Supplementary Material online). In the WES dataset with only North Africans, 7 random individuals were selected per population. In all cases, 3 replicates were performed with similar results.

Variant Annotation

Variant deleteriousness was assessed based on GERP RS scores collected from the CADD online tool. GERP RS scores are based on the predicted effect of allele substitutions based on sequence conservation (Cooper et al. 2005; Davydov et al. 2010). Categorization of GERP RS scores in different levels of predicted deleteriousness was done in the same way as in Henn et al. (2015) and Henn et al. (2016). Two additional methods to predict deleteriousness were used for independent comparisons: PolyPhen-2 (Adzhubei et al. 2010) and CADD scores (Kircher et al. 2014; Rentzsch et al. 2019). PolyPhen-2 scores were collected from the Ensembl Variant Effect Predictor (McLaren et al. 2016) and categorized according to the proposed categories in the Ensembl Variant Effect Predictor. Variants labeled as "unknown" were left out for the analyses using PolyPhen-2 scores. CADD scores were collected from the CADD online tool and categorized following the recommendations in the CADD online site (https://cadd.gs.washington.edu/info) and the Ensembl Variant Effect Predictor (McLaren et al. 2016).

The genomic effect of variants (i.e. synonymous or missense) was annotated using the Ensembl Variant Effect Predictor (McLaren et al. 2016).

ROH Analysis

ROHs were detected in the complete WES dataset and in the merged dataset, using the same approach. ROHs were detected at the individual level with PLINK 2.0 (Purcell and Chang; Chang et al. 2015) following the WES-optimized parameters used in Nutile et al. (2019) and Lucas-Sánchez, Font-Porterias, et al. (2021). Pruning kept 309,475 and 627,772 variants in the WES and merged datasets, respectively. Per-individual total number of ROHs (Fig. 2a and supplementary fig. S10, Supplementary Material online), per-individual total length (sum of all ROH lengths) (supplementary figs. S8 and S11, Supplementary Material online), and per-individual average ROH length (supplementary figs. S9 and S12, Supplementary Material online) were calculated in each population. The total number and total length of ROHs

were calculated in 3 different ROH length categories, accounting for the different demographic interpretations of ROHs depending on their length (Ceballos et al. 2018): 1 to 2.5, 2.5 to 5, and >5 Mb. The statistical significance of the differences between population pairs of interest was tested with *t*-tests.

A second group of ROH analyses consisted of assessing the ratio of missense homozygous derived genotypes to synonymous homozygous derived genotypes in 5 different exomic regions: inside and outside ROH tracts and inside regions occupied by ROHs in each of the 3 length categories (supplementary figs. S22 to S24, Supplementary Material online). In each region, ratio calculations were divided per GERP RS score category, selecting only missense deleterious sites in the corresponding GERP RS score range for the 3 categories that predict SNPs to be deleterious ($2 \leq$ GERP RS score < 4, $4 \leq$ GERP RS score < 6, and GERP RS score ≥ 6). VCFtools 0.1.14 (Danecek et al. 2011) was used to calculate individual allele counts to later select homozygous sites. Ratios were calculated per individual and *t*-tests were used to assess statistical significance of the differences between regions within each population and GERP RS score category.

SFS and Distribution of Variants

Different unfolded SFSs were computed using the ancestral-state-annotated versions of the 2 WES datasets. VCFtools 0.1.14 (Danecek et al. 2011) was used for allele count calculation in each population, with which the frequencies at each derived allele count bin were calculated after removing the 2 fixed categories. Population sizes were thinned to match the size of the population with less samples, randomly selecting the kept individuals. Different iterations were performed to confirm that selecting different individuals did not significantly change the results.

For each dataset, 4 different groups of SFS were performed: (i) using all sites (Fig. 3a and supplementary fig. 14a, Supplementary Material online) and dividing sites by (ii) GERP RS score (Fig. 3b to e and supplementary fig. 14b to e, Supplementary Material online), (iii) PolyPhen-2 (supplementary fig. S17, Supplementary Material online), and (iv) CADD score (supplementary fig. S15, Supplementary Material online) categories.

To complement the SFS analysis, sites with at least 1 derived allele and with a predicted deleterious GERP RS score (GERP ≥ 2) were grouped and divided into low frequency (singletons and doubletons) and common sites (from tripletons to higher frequencies) (Fig. 3f). These groups were compared between populations, and confidence intervals were calculated with the Wald method. The same was performed using PolyPhen-2 (supplementary fig. S18, Supplementary Material online) and CADD (supplementary fig. S16, Supplementary Material online) scores instead of GERP RS scores.

Genetic Load

The current genetic load was calculated as the mutational load, this is, the genetic load measured as the accumulation of deleterious mutations. Differences in mutational load between population pairs were assessed by computing 2 commonly used summary statistics based on individual genotypic data: per-individual number of derived alleles (N_{alleles}) and per-individual homozygous derived genotypes (N_{hom}). These statistics were calculated per population and then population-pairs ratios were computed (Fig. 4). This was performed by selecting derived sites according to their GERP RS score and dividing the analysis into the 4 GERP RS score categories suggested by Henn et al. (2015) and Henn et al. (2016). The number of alleles and homozygous genotypes was calculated directly from the VCF file with VCFtools 0.1.14 (Danecek et al. 2011). Bootstrap reanalysis with 1,000 iterations was performed to compute confidence intervals and assess statistical significance. The variant set was divided into 1,000 blocks and 1,000 random blocks were taken in each iteration allowing resampling, which takes into account demographic variance. Quantiles 0.025 and 0.975 of the bootstrap distribution were set as the confidence intervals for each ratio. We followed Lopez et al. (2018) and required $P < 0.001$ to declare significance. The same analysis was repeated taking into account the synonymous and missense labeling from the Ensembl Variant Effect Predictor (McLaren et al. 2016) and thus selecting only variants labeled as synonymous for those with GERP < 2 and variants labeled as missense for those with GERP ≥ 2 (supplementary fig. S19, Supplementary Material online). Additional replicates of the analysis were performed using PolyPhen-2 (supplementary fig. S21, Supplementary Material online) and CADD (supplementary fig. S20, Supplementary Material online) scores instead of GERP RS scores.

DFE of New Deleterious Mutations

DFE was calculated per population using the ∂a∂i (Gutenkunst et al. 2009) and Fit∂a∂i (Kim et al. 2017) methods. Due to the absence of a refined demographic model for any North African population, 5 different general models were tested using ∂a∂i: the standard neutral model, a growth model, a bottle growth model, a 2-epoch model, and a 3-epoch model (https://dadi.readthedocs.io/en/latest/api/dadi/Demographics1D.html). The input for these analyses was the unfolded synonymous SFS, which serves as a proxy for the neutral variation of the population. For each population, the model with the best likelihood was kept for the subsequent steps. The selected models and estimated demographic parameters were then used in Fit∂a∂i to fit the data to a gamma distribution, resulting in a DFE per population with shape and scale parameters (supplementary fig. S27, Supplementary Material online). For parameters and model likelihoods inferred with ∂a∂i and Fit∂a∂i, confidence intervals were calculated as the 0.025 and 0.975 quantiles of a

bootstrap-by-site distribution as described in the section above. The goodness of fit for the selected best-likelihood model was assessed after ∂a∂i (supplementary fig. S25, Supplementary Material online) and Fit∂a∂i (supplementary fig. S26, Supplementary Material online) analyses with χ^2 tests comparing the expected synonymous (∂a∂i) or missense (Fit∂a∂i) SFS according to the model and the observed synonymous (∂a∂i) or missense (Fit∂a∂i) SFS.

Estimation of Demographic Parameters

For the subsequent steps, we required the mean parameter of the DFE gamma distributions, i.e. the $E(s)$. Fit∂a∂i outputs the $E(N_e s)$, which is shape × scale, so we calculated $E(s) = E(N_e s)/N_w$, where N_w is the effective population size weighted along time (Lopez et al. 2018). As we only required these parameters in a subset of populations from our dataset, which all have 3-epoch models as the best-likelihood model, N_w was calculated as $N_1 w_1 + N_2 w_2 + N_3 w_3$, where N_i is the effective population size in epoch i and w_i is an assigned weight to generation i, as explained in Lopez et al. (2018). We calculated $N_1 = N_{ANC}$ from $\theta s = 4N_{ANC}\mu L_S$ (Kim et al. 2017), where θs is the population-scaled synonymous mutation rate estimated with ∂a∂i, and $\mu = 1.5 \times 10^{-8}$ (Ségurel et al. 2014). L_S comes from $L = L_{NS} + L_S$, being L the number of bases from which variants that entered the analyses were called, reducing the initial called variants by the fraction of variants eliminated by filtering (in our case, 37,000,000 variants), and assuming 2.31 as the nonsynonymous-to-synonymous sites ratio (L_{NS}/L_S) (Huber et al. 2017; Kim et al. 2017). N_2 and N_3 were calculated from ∂a∂i estimated parameters following the software documentation and the developer's recommendations. Thus, $N_2 = \text{nuB} * N_{ANC}$, and $N_3 = \text{nuF} * N_{ANC}$, where nuB and nuF are the ∂a∂i estimated effective population size ratios for the second and third epochs of the model. As explained in Lopez et al. (2018), weight calculations require a time parameter for the second and third epochs, which we also calculated from ∂a∂i estimates as follows: $t_2 = \text{TB} * 2N_{ANC}$, and $t_3 = \text{TF} * 2N_{ANC}$, being TB the length of the bottleneck and TF the time since bottleneck recovery, both in units of $2N_{ANC}$ generations.

Trajectory of Genetic Load

The trajectory of the genetic load was assessed from forward simulations performed with the software SLiM 3.7.1 (Haller and Messer 2019). Due to the absence of a known demographic model, the IBDNe trajectories were used as proxies. For each population, population size values were taken from the IBDNe trajectory every 10 generations starting at 100 generations ago. To reduce computational power requirements, we used the tree-sequence recording approach as described in SLiM's manual. The recombination rate was set to 10^{-8} per base per generation, following Lopez et al. (2018) and Font-Porterias et al. (2021). Because of the use of tree-sequencing recording, the mutation rate was adjusted from an initial value of 1.36×10^{-8} per base position per generation (Lopez et al. 2018; Font-Porterias et al. 2021), to 8.37×10^{-11}, which is the fraction corresponding to nonneutral mutations. Different numbers of generations for the burn-in phase were tested with similar results. We simulated a genomic structure consisting of 20 unlinked chromosomes with 1,000 genes separated by 50,000-base-long neutral non-coding regions. Genes were divided into 8 exon–intron pairs, with the exons being 100-base long and introns 5,000-base long. Introns were assumed to be neutral, and exons were formed by 3-base pair codons with the first 2 positions under selection and accepting deleterious mutations and the third being neutral (Lopez et al. 2018; Font-Porterias et al. 2021). Deleterious mutations were set to follow a DFE with a gamma distribution with mean $E(s)$ and shape parameters corresponding to each population's estimates in the DFE and demographic parameters estimation analyses. After the burn-in phase (5,000 generations), nonfixed mutations were sampled from the simulated population every 5 generations, and the mutational load was calculated as $L = 1 - \exp(\sum_i l_i)$, where i is each mutation and $l = s \times (2hq + (1-2h) \times q^2)$, being s the selection coefficient, h the dominance coefficient, and q the frequency of the mutation. For each population, 2 scenarios were simulated: a fully additive model ($h = 0.5$) and a fully recessive model ($h = 0$).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank all the volunteers involved in this study.

Author Contributions

M.L.-S. and D.C. designed the study. M.L.-S. conducted the analysis. M.L.-S., F.C., and D.C. contributed to the interpretation of the data. M.L.-S. wrote the manuscript with the help of D.C. A.A., A.B., and T.B. contributed to the sampling and helped contextualize the results. All authors revised and approved the manuscript.

Funding

This work was supported by the Spanish Ministry of Science and Innovation (grant numbers I-COOP0018, PID2019-106485GB-I00, and PID2022-138755NB-I00) and "Unidad María de Maeztu" (CEX2018-000792-M) funded by the MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

Conflict of interest statement. None declared.

Data Availability

Algerian Amazigh and non-Amazigh whole-exome sequences are deposited at EGA accession number: EGAS00001007236. Algerian Amazigh and non-Amazigh genotypes are deposited at EGA accession number: EGAS00001007235.

References

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015:**526**(7571):68–74. https://doi.org/10.1038/nature15393.

Abdeli A. 2021. Variations géographiques de polymorphismes génétiques dans la population Algérienne (PhD thesis). Bab Ezzouar: Faculté de Chimie, Université des Sciences et de la Technologie Houari Boumediene.

Adams AM, Hudson RR. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*. 2004:**168**(3):1699–1712. https://doi.org/10.1534/genetics.104.030171.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010:**7**(4):248–249. https://doi.org/10.1038/nmeth0410-248.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009:**19**(9):1655–1664. https://doi.org/10.1101/gr.094052.109.

Alport EA. The Mzab. *J R Anthropol Inst Gt Britain Irel*. 1954:**84**:34.

Anagnostou P, Dominici V, Battaggia C, Boukhchim N, Ben Nasr J, Boussoffara R, Cancellieri E, Marnaoui M, Marzouki M, Bel Haj Brahim H, *et al*. Berbers and Arabs: tracing the genetic diversity and history of Southern Tunisia through genome wide analysis. *Am J Phys Anthropol*. 2020:**173**(4):697–708. https://doi.org/10.1002/ajpa.24139.

Anwar WA, Khyatti M, Hemminki K. Consanguinity and genetic diseases in North Africa and immigrants to Europe. *Eur J Public Health*. 2014:**24**(suppl 1):57–63. https://doi.org/10.1093/eurpub/cku104.

Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, Fadhlaoui-Zid K, Zalloua P, Hellenthal G, Comas D. Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Mol Biol Evol*. 2017:**34**(2):318–329. https://doi.org/10.1093/molbev/msw218.

Bachir S, Aouar A. Study of the impact of consanguinity on abortion and mortality in the population of Beni Abbes (soutwestern Algeria). *Egypt J Med Hum Genet*. 2019:**20**(1):1. https://doi.org/10.1186/s43042-019-0004-7.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*. 2016:**32**(18):2817–2823. https://doi.org/10.1093/bioinformatics/btw327.

Bekada A. *Caractérisation anthropogénétique d'un échantillon de la population algérienne: analyse des marqueurs parentaux*. Oran: Faculté des Sciences de la Nature et de la Vie, Univesité d'Oran; 2015.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014:**30**(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bosch AE, Calafell F, Comas D, Mateu E. Population history of North Africa: evidence from classical genetic markers. *Hum Biol*. 1997:**69**:295–311.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, *et al*. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. 2008:**4**(5):e1000083. https://doi.org/10.1371/journal.pgen.1000083.

Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013:**194**(2):459–471. https://doi.org/10.1534/genetics.113.150029.

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**(3):404–418. https://doi.org/10.1016/j.ajhg.2015.07.012.

Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet*. 2021:**108**(10):1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005.

Camps G. *Les berbères: mémoire et identité*. Arles: Errance; 1995.

Camps G. *Los bereberes, ¿mito o realidad?* Barcelona: Icaria; 1996.

Camps G. *Los bereberes: de la orilla del mediterráneo al límite meridional del Sáhara*. Barcelona: Icaria; 1998.

Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen JB, Frazer KA. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Nephrol*. 2014:**15**:125. https://doi.org/10.1186/1471-2369-15-1.

Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, Grenier J-CC, Gbeha E, Hamdan FF, Girard S, *et al*. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet*. 2013:**9**(9):e1003815. https://doi.org/10.1371/journal.pgen.1003815.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet*. 2018:**19**(4):220–234. https://doi.org/10.1038/nrg.2017.109.

Chang CC, Chow CC, Tellier LCAM, 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.

Cherni L, Goios A, Yacoubi B, Benammar A, Slama A. Y-chromosomal STR haplotypes in three ethnic groups and one cosmopolitan population from Tunisia. *Forensic Sci Int*. 2005:**152**(1):95–99. https://doi.org/10.1016/j.forsciint.2005.02.007.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005:**15**(7):901–913. https://doi.org/10.1101/gr.3577405.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al*. The variant call format and VCFtools. *Bioinformatics*. 2011:**27**(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. 2010:**6**(12):e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet*. 2015:**47**(2):126–131. https://doi.org/10.1038/ng.3186.

Fadhlaoui-Zid K, Khodjet-el-khil H, Mendizabal I, Benammar-elgaaied A, Comas D. Genetic structure of Tunisian ethnic groups revealed by paternal lineages. *Am J Phys Anthropol*. 2011:**146**(2):271–280. https://doi.org/10.1002/ajpa.21581.

Fadhlaoui-Zid K, Plaza S, Calafell F, Ben AM, Comas D, El gaaied AB. Mitochondrial DNA heterogeneity in Tunisian Berbers. *Ann Hum Genet*. 2004:**68**(3):222–233. https://doi.org/10.1046/j.1529-8817.2004.00096.x.

Fadhlaoui-Zid K, Rodríguez-Botigué L, Naoui N, Benammar-Elgaaied A, Calafell F, Comas D. Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley. *Am J Phys Anthropol*. 2011:**145**(1):107–117. https://doi.org/10.1002/ajpa.21472.

Font-Porterias N, Caro-Consuegra R, Lucas-Sánchez M, Lopez M, Giménez A, Carballo-Mesa A, Bosch E, Calafell F, Quintana-Murci L, Comas D. The counteracting effects of demography on functional genomic variation: the Roma Paradigm.

Mol Biol Evol. 2021:**38**(7):2804–2817. https://doi.org/10.1093/molbev/msab070.

Fregel R, Méndez FL, Bokbot Y, Martín-Socas D, Camalich-Massieu MD, Santana J, Morales J, Ávila-Arcos MC, Underhill PA, Shapiro B, *et al*. Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. *Proc Natl Acad Sci U S A*. 2018:**115**(26):6774–6779. https://doi.org/10.1073/pnas.1800851115.

Fu W, Gittelman RM, Bamshad MJ, Akey JM. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am J Hum Genet*. 2014:**95**(4):421–436. https://doi.org/10.1016/j.ajhg.2014.09.006.

Ghaki M. Els Berbers. In: *Tunísia, terra de cultures. Tunisia, land of cultures*. Barcelona: IEMed-MuPCVa; 2003. p. 39–42.

Gravel S. When is selection effective? *Genetics*. 2016:**203**(1):451–462. https://doi.org/10.1534/genetics.115.184630.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 2009:**5**(10):1000695. https://doi.org/10.1371/journal.pgen.1000695.

Haber M, Saif-Ali R, Alhabori M, Chen Y, Platt DE, Tyler-Smith C, Xue Y. Insight into the genomic history of the near east from whole-genome sequences and genotypes of Yemenis. *bioRxiv*. 2019. https://doi.org/10.1101/749341.

Haldane J. A mathematical theory of natural and artificial selection, part V: selection and mutation. *Math Proc Cambridge Philos Soc*. 1927:**23**(7):838–844. https://doi.org/10.1017/S0305004100015644.

Haldane J. The effect of variation of fitness. *Am Nat*. 1937:**71**(735):337–349. https://doi.org/10.1086/280722.

Haller BC, Messer PW. SLim 3: forward genetic simulations beyond the Wright–Fisher model. *Mol Biol Evol*. 2019:**36**(3):632–637. https://doi.org/10.1093/molbev/msy228.

Harich N, Costa MD, Fernandes V, Kandil M, Pereira JB, Silva NM, Pereira L. The trans-Saharan slave trade—clues from interpolation analyses and high-resolution characterization of mitochondrial DNA lineages. *BMC Evol Biol*. 2010:**10**(1):138. https://doi.org/10.1186/1471-2148-10-138.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. *Nat Rev Gene*. 2015:**16**(6):333–343. https://doi.org/10.1038/nrg3931.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, *et al*. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet*. 2012:**8**(1):e1002397. https://doi.org/10.1371/journal.pgen.1002397.

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, *et al*. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U S A*. 2016:**113**(4):E440-E449. https://doi.org/10.1073/pnas.1510805112.

Hiernaux J. *The people of Africa*. New York: Encore Editions; 1975.

Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc Natl Acad Sci U S A*. 2017:**114**(17):4465–4470. https://doi.org/10.1073/pnas.1619508114.

Ibn-Khaldoun A. *Histoire des berberes et des dynasties muslmanes de l'Afrique Septentrionale: traduction de Le Baronde Slane*. Paris: Paul Geuthner; 1968.

Idaghdour Y, Czika W, Shianna K V, Lee SH, Visscher PM, Martin HC, Miclaus K, Jadallah SJ, Goldstein DB, Wolfinger RD, *et al*. Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat Genet*. 2010:**42**(1):62–67. https://doi.org/10.1038/ng.495.

Kim BY, Huber CD, Lohmueller KE. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*. 2017:**206**(1):345–361. https://doi.org/10.1534/genetics.116.197145.

Kimble GHT. The Berbers of Eastern Algeria. *Geogr J*. 1941:**97**(6):337. https://doi.org/10.2307/1788169.

Kimura M, Maruyama T, Crow JF. The mutation load in small populations. *Genetics*. 1963:**48**(10):1303–1312. https://doi.org/10.1093/genetics/48.10.1303.

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**(3):310–315. https://doi.org/10.1038/ng.2892.

Kousathanas A, Oliver F, Halligan DL, Keightley PD. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol Biol Evol*. 2011:**28**(3):1183–1191. https://doi.org/10.1093/molbev/msq299.

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009:**25**(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, Esko T, Mägi R, Inouye M, Lappalainen T, *et al*. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet*. 2014:**10**(7):e1004494. https://doi.org/10.1371/journal.pgen.1004494.

Lohmueller KE. The distribution of deleterious genetic variation in human populations. *Curr Opin Genet Dev*. 2014:**29**:139–146. https://doi.org/10.1016/j.gde.2014.09.005.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, *et al*. Proportionally more deleterious genetic variation in European than in African populations. *Nature*. 2008:**451**(7181):994–997. https://doi.org/10.1038/nature06611.

Lopez M, Kousathanas A, Quach H, Harmant C, Mouguiama-Daouda P, Hombert JM, Froment A, Perry GH, Barreiro LB, Verdu P, *et al*. The demographic history and mutational load of African hunter-gatherers and farmers. *Nat Ecol Evol*. 2018:**2**(4):721–730. https://doi.org/10.1038/s41559-018-0496-4.

Lucas-Sánchez M, Fadhlaoui-Zid K, Comas D. The genomic analysis of current-day North African populations reveals the existence of trans-Saharan migrations with different origins and dates. *Hum Genet*. 2023:**142**(2):305–320. https://doi.org/10.1007/s00439-022-02503-3.

Lucas-Sánchez M, Font-Porterias N, Calafell F, Fadhlaoui-Zid K, Comas D. Whole-exome analysis in Tunisian Imazighen and Arabs shows the impact of demography in functional variation. *Sci Rep*. 2021:**11**(1):21125. https://doi.org/10.1038/s41598-021-00576-0.

Lucas-Sánchez M, Serradell JM, Comas D. Population history of North Africa based on modern and ancient genomes. *Hum Mol Genet*. 2021:**30**(R1):R17–R23. https://doi.org/10.1093/hmg/ddaa261.

Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*. 2004:**166**(1):351–372. https://doi.org/10.1534/genetics.166.1.351.

Masonen P. Trans-saharan trade and the West African discovery of the Mediterranean world. In: Sabour M, Vikør KS, editors. *Ethnic encounter and culture change*. Bergen: C. Hurst & Co.; 1997. p. 116–142.

McEvedy C. *The Penguin Atlas of African history*. London: Penguin Books; 1995.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010:**20**(9):1297–1303. https://doi.org/10.1101/gr.107524.110.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. *Genome Biol*. 2016:**17**(1):122. https://doi.org/10.1186/s13059-016-0974-4.

Mezzavilla M. Neon: an R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPS. *J Comput Sci Syst Biol*. 2015:**8**(1):37–44. https://doi.org/10.4172/jcsb.1000168.

Morton NE, Crow JF, Muller HJ. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc Natl Acad Sci U S A*. 1956:**42**(11):855–863. https://doi.org/10.1073/pnas.42.11.855.

Muller HJ. Our load of mutations. *Am J Hum Genet*. 1950:**2**:111–176.

Newman JL. *The peopling of Africa: a geographic interpretation*. New Haven: Yale University Press; 1995.

Nutile T, Ruggiero D, Herzig AF, Tirozzi A, Nappo S, Sorice R, Marangio F, Bellenguez C, Leutenegger AL, Ciullo M. Whole-exome sequencing in the isolated populations of Cilento from South Italy. *Sci Rep*. 2019:**9**(1):1–13. https://doi.org/10.1038/s41598-019-41022-6.

Office National des Statistiques. "Résultats Du Recensement Général De La Population et De L'Habitat 2008 (Ménager Ordiniaires et Collectifs)"; 2009 [Accessed 2023 Sept 2]. https://www.ons.dz/collections.

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

Pedersen CET, Lohmueller KE, Grarup N, Bjerregaard P, Hansen T, Siegismund HR, Moltke I, Albrechtsen A. The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: insights from the Greenlandic Inuit. *Genetics*. 2017:**205**(2):787–801. https://doi.org/10.1534/genetics.116.193821.

Pellat Ch, Yver G, Basset R, Galand L. Berbers. In: *Encyclopaedia of Islam*. Leiden: Brill Academic Publishers, 2nd ed. 2012. http://dx.doi.org/10.1163/1573-3912_islam_COM_0114.

Purcell S, Chang C. PLINK 1.9. Available from: https://www.cog-genomics.org/plink/.

Purcell S, Chang C. PLINK 2.0. Available from: https://www.cog-genomics.org/plink/2.0/.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019:**47**(D1):D886–D894. https://doi.org/10.1093/nar/gky1016.

Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*. 2014:**15**(1):47–70. https://doi.org/10.1146/annurev-genom-031714-125740.

Serra-Vidal G, Lucas-Sánchez M, Fadhlaoui-Zid K, Bekada A, Zalloua P, Comas D. Heterogeneity in Palaeolithic population continuity and Neolithic expansion in North Africa. *Curr Biol*. 2019:**29**(22):3953–3959.e4. https://doi.org/10.1016/j.cub.2019.09.050.

Simões LG, Günther T, Martínez-Sánchez RM, Vera-Rodríguez JC, Iriate E, Rodríguez-Varela R, Bokbot Y, Valdiosera C, Jakobsson M. Northwest African Neolithic initiated by migrants from Iberia and Levant. *Nature*. 2023:**618**(7965):550–556. https://doi.org/10.1038/s41586-023-06166-6.

Simons YB, Sella G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Curr Opin Genet Dev*. 2016:**41**:150–158. https://doi.org/10.1016/j.gde.2016.09.006.

Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. *Nat Genet*. 2014:**46**(3):220–224. https://doi.org/10.1038/ng.2896.

Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, Li JZ. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet*. 2013:**93**(1):90–102. https://doi.org/10.1016/j.ajhg.2013.05.003.

Tremblay M, Vézina H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am J Hum Genet*. 2000:**66**(2):651–658. https://doi.org/10.1086/302770.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, *et al.* From fastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013:**43**(1):11.10.1–11.10-33. https://doi.org/10.1002/0471250953.bi1110s43.

Wright S. The distribution of gene frequencies in populations. *Science*. 1937:**85**(2212):504. https://doi.org/10.1126/science.85.2212.504.a.

Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. 2014:**30**(7):1006–1007. https://doi.org/10.1093/bioinformatics/btt730.