Title: Gene flow from North Africa contributes to differential human genetic diversity in Southern Europe

Authors: Laura R. Botigué1*, Brenna M. Henn2*, Simon Gravel2, Brian K. Maples2, Christopher R. Gignoux3, Erik Corona4,5, Gil Atzmon6,7, Edward Burns6, Harry Ostrer7,8, Carlos Flores9,10, Jaume Bertranpetit1, David Comas1, David Bustamante2

Affiliation:
1 Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.
2 Department of Genetics, Stanford University, Stanford, CA, USA.
3 University of California San Francisco, San Francisco, CA, USA.
4 Department of Pediatrics, Stanford University, Stanford, CA, USA.
5 Lucile Packard Children’s Hospital, Stanford CA, USA.
6 Department of Medicine, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
7 Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
8 Department of Pathology, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
9 Research Unit, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain
10 CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain
* Both authors contributed equally
^ Current address: Department of Ecology and Evolution, Stony Brook University, NY 11794, USA.
□ Co-corresponding authors and joint supervisors

Word Count: 5,250
Running Title: Gene flow from North Africa to Europe
Abstract:

Human genetic diversity in southern Europe is higher than in other regions of the continent. This difference has been attributed to post-glacial expansions, the demic diffusion of agriculture from the Near East and gene flow from Africa. Using single nucleotide polymorphism (SNP) data from 2,099 individuals in 43 populations, we show that estimates of recent shared ancestry between Europe and Africa are substantially increased when gene flow from North Africans, rather than Sub-Saharan Africans, is considered. The gradient of North African ancestry accounts for previous observations of low levels of sharing with Sub-Saharan Africa and is independent of recent gene flow from the Near East. The source of genetic diversity in southern Europe has important biomedical implications; we find that most disease risk alleles from genome-wide association studies follow expected patterns of divergence between Europe and North Africa, with the principal exception of multiple sclerosis.

Introduction

Clinal gradients of human genetic diversity and genetic markers in Europe have been attributed to directional migration patterns, climate, natural selection and isolation by distance models (1-4). In particular, southern European populations exhibit the highest levels of genetic diversity, which declines in northern latitudes. Three main hypotheses have been proposed to explain this phenomenon. Under the first hypothesis, populations retreated to glacial refugia in southern Europe about 20,000 years ago (ya), but when these populations later re-colonized the continent, only a subset of the genetic diversity was carried into northern regions (5). The second hypothesis is that gene flow from the Near East, associated with the demic diffusion of agriculture, differentially affected geographic regions and in particular introduced additional genetic diversity to southeastern Europe (6, 7). The third hypothesis suggests that increased genetic diversity is the result of migrations from the African continent into southern Europe (8, 9). These hypotheses are not mutually exclusive; however, we focus on testing a hypothesis of gene flow from Africa to Europe, which has received the least amount of attention and may be the easiest to detect due to the recent timeframe of the proposed demographic event.

About 20,000 ya during the Last Glacial Maximum, populations in Europe retreated into the glacial refugia located in the Mediterranean peninsulas, where climate conditions were milder. Differences in genetic diversity in extant European populations have been explained by a re-colonization from these glacial refugia at the end of the glacial period, a process during which only a subset of the genetic diversity from the refugia would expand into the rest of the continent. For instance, radiocarbon dates suggest that re-colonization of Britain took place around 14,700 ya (10). The geographic distribution and ages of mtDNA haplogroup HV0, V, H1 and H3 in European populations reflect that pattern of postglacial human re-colonization from the Franco-Cantabrian refugia (11-13), and a similar pattern has also been detected in Y chromosome as the case of haplogroup I (14). Differential gradients of genetic diversity in many other species within Europe (e.g. grasshoppers, brown bears and oak trees) have also been attributed to post-glacial expansions during this time (15).

Changes in genetic diversity in European populations have also been associated with the Neolithic expansion from the Near East (7). The relative effect of demic
diffusion of early agriculture on the genetic composition of European populations remains a hotly contested topic (16-18). It has been suggested that Near Eastern Neolithic mtDNA lineages comprise almost one quarter of the extant European haplogroups (19) and Y chromosome genetic diversity also retains a strong signal from the Near East (20). Beginning about 8,000 years ago, extensive archaeological data document the spread of the Neolithic across southern Europe; for example, at this time similar Neolithic pottery is found in both Europe and the Near East. A notable exception is southern Iberia: around 7,500 ya, strong similarities are found in pottery production in this region and Northwest Africa. Additionally, the existence of “maritime pioneers” in the Mediterranean Sea during this period has been hypothesized (21). Some authors have interpreted this as a proof of the existence of Neolithic networks joining the European and African shores of western Mediterranean Sea (22).

Lastly, three recent studies highlight the possibility of genetic exchange between Europe and Africa. Moorjani et al. estimated that about 1-3% of recent Sub-Saharan African ancestry is present in multiple southern European populations, Cerezo et al. find evidence of older (11,000 ya) Sub-Saharan gene flow towards Europe based on mtDNA genomes, and Auton et al. found that short haplotypes were shared between the Yoruban Nigerians and southwestern Europeans. However, given the geographic barrier imposed by the Sahara Desert between North Africa and Sub-Saharan Africa, and the proximity of North Africa to Europe, it is plausible that gene flow from Africa to Europe actually originated in North Africa. North Africans are significantly genetically diverged from Sub-Saharan populations (24, 25), and hence previous studies may not have accurately estimated the proportion or range of admixture in Europe by using a Sub-Saharan sample as a source population. For example, the Moorish Berber conquest in Iberia began in the 8th century C.E. and lasted for more than five hundred years; this conquest has been suggested as a potential source of gene flow from North Africa towards the Iberian Peninsula. The Y chromosome haplogroup E3b2-M81 distribution is in agreement with a recent North African gene flow at that period (26).

Here we analyze recently published SNP data from 7 North African populations (25), together with data from 30 European populations (25, 27) (including new Affymetrix 6.0 data for 3 Spanish populations: Galician, Andalucian, Canary Islands), 2 European Jewish populations (28), 1 Near Eastern population (29) and HapMap3 Sub-Saharan African populations (Table S1). Our objective was to quantify the extent and pattern of recent gene flow between European and African populations. We use both allele frequency and haplotype identity by descent (IBD) approaches to estimate North African ancestry proportions in European populations. In order to quantify the variance in ancestry in European populations and obtain bounds on the time since admixture, we use a quantitative model for the decrease in ancestry variance with the time since admixture (30). Recent gene flow between populations can also be detected by long haplotypes shared IBD with high density SNP genotyping data (31, 32). We investigate regional patterns of haplotype sharing between North Africa, Sub-Saharan Africa, Near East and Europe in detail, and observe a significant latitudinal gradient of North African ancestry within Europe characterized by a dramatic difference between the Iberian Peninsula and the neighboring regions.

Results:
Estimating Gene Flow between Africa and Europe

Ancestry proportions:

To estimate allele-based sharing between Africans and Europeans, we applied an unsupervised clustering algorithm, ADMIXTURE (33), to data from all populations (Table S1). We assumed $k=2$-10 ancestral populations and performed 10 iterations for each $k$ (Fig. S1, S2). Importantly, we make no assumption in this analysis that any given source population is an un-admixed population; that is, the analysis is run unsupervised where the current population configurations are not known (such that any sub-Saharan African ancestry could be detected in both North Africans and Europeans). Estimates of admixture show little bias using an unsupervised approach when the ancestral populations are substantially diverged (e.g. at an $F_{st}$ of 0.05) and using several hundred thousand markers (34); previous work suggests that Europeans and North Africans are segregated by an $F_{st}$=0.06 (25). As the number of $k$ ancestral clusters increased, we observed several population-specific ancestry clusters emerge supporting the presentation of higher $k$ ancestries. We conservatively present $k=3$ through 6 (Fig. 1) but additional results are presented in the Supplement.

At $k$=4, the ancestry assignment differentiated between non-Jewish European populations (from now on referred to as “European”), European Jews, Sub-Saharan Africans, and a group formed by Near Eastern and North African populations. At $k=5,6$ components mainly assigned to North African populations and Tunisian Berbers, respectively, clearly appear. European populations sharing this North African ancestral component are almost exclusively in southern Europe (Fig. 1, Fig S3). Southern European populations have a high proportion (5-35%) of joint Near Eastern $|$ North African ancestry assigned at $k$=4. However, identification of distinct Near Eastern and North African ancestries in $k\geq5$ differentiates southeastern from southwestern Europe. Southwestern European populations average between 4-20% of their genomes assigned to North African ancestral cluster (Fig. S3), while this value does not exceed 2% in southeastern European populations. Contrary to past observations, Sub-Saharan ancestry is detected at <1% in Europe, with the exception of the Canary Islands. In summary, when North African populations are included as a source, allele frequency-based clustering indicates better assignment to North African than to Sub-Saharan ancestry, and estimates of African ancestry in European populations increase relative to previous studies. European ancestry is also detected in North African populations. At $k=6$ it ranges between 4 – 16% in the rest of North Africa, with notable intra-population variation (35) and is absent in most Magrebi (western North African) individuals from Tunisia and Western Sahara.

In order to test whether our results were robust to the inference procedure in ADMIXTURE, we additionally employ a non-Bayesian approach to estimate North African ancestry in southern Europeans. With the program RFMix (36) we assumed that there are three putative source populations for ancestry in Europeans: German, Saharawi and Qatari. Estimates of North African ancestry range between 5-14% in the European populations and trends of the overall ancestry clines are concordant with ADMIXTURE (Table S4, Fig. S15). We tested whether ADMIXTURE could accurately infer North African ancestry proportions in Europeans via simulation of historical admixture
scenarios; we find that $k=4,5$ gave more accurate admixture estimates of North African ancestry. The correlation between the simulated North African ancestry and the one inferred with ADMIXTURE dramatically increases from $k=3$ to 4 in all simulated populations (Figure S15) and the average difference in ancestry proportions at the individual level decreases from 0.04 to 0.02 when 4 or 5 ancestral components are considered.

**Isolation By Distance:**
We note that ADMIXTURE analysis may lead to artifactual ancestral components when the populations tested follow an isolation by distance model (37). In order to determine whether the North African component detected by ADMIXTURE reflects admixture from distinct source populations or is a consequence of an isolation by distance process, we performed a Mantel test comparing pairwise genetic and geographic distances among European and North African populations. The great circle geographic distances between populations were calculated including a western waypoint located at the Gibraltar Strait for North Africa, following (38). The Mantel test was performed using the software IBWS v3.23 (39). When all European and North African populations are included in the analysis there is a positive correlation between genetic and geographic distances of $r^2=0.268$. However, this result is driven by European vs. European pairwise comparisons (SI Appendix). When we compared genetic and geographic distances focusing only on pairwise European vs. North African comparisons, no correlation between genetic and geographic distance is found, $r^2 < 0.001$ ($p = 0.931$), ruling out the hypothesis that gene flow between North Africa and Europe follows an isolation by distance model (Fig. S14).

**Long identical by descent haplotypes:**
Recent gene flow between populations results in long shared haplotypes. We investigated differences in African ancestry among European populations by analyzing these genomic segments inferred to be identical by descent between Sub-Saharan Africa, North Africa, Europe and the Near East (SI Appendix). Migration from one endogamous population to another generates genetic segments that share a recent common ancestor (and in short time spans are IBD) between the two populations; the distribution and length of IBD segments are informative of recent migration. We restrict our analysis to IBD segments greater than 1.5 cM identified using fastIBD (40). Long IBD segments are far less sensitive to ascertainment bias than are allele frequency estimates, and by analyzing segments >1.5 cM, we help minimize background linkage disequilibrium (problematic for the identification of short haplotypes) (41). We first examined the extent to which detection of long IBD segments is conditioned on marker density. We compared IBD performance between a high-density dataset (HDD) of 641,884 SNPs and a low-density dataset (LDD) of 280,462 SNPs from our primary analysis. We found that when considering IBD sharing between North Africa and Europe, 72% of the segments found in LDD overlap with those found in HDD. However, when only IBD sharing within Europe is considered, this overlap increases to 80% (SI Appendix); the estimates here are similar to the power estimates in (40) for fastIBD using the same parameters.

We calculated a summary statistic informative of the level of gene flow between two populations, though not the directionality: “$W_{EA}$” is the sum of lengths (in cM) of all DNA segments inferred to be shared identical by descent between a given European population “$E$” and North African | Sub-Saharan African populations “$A$”, normalized by
the average sample size and scaled here by 100 (28). We note that extensive IBD sharing may be a signal of positive selection shared among populations (42). To confirm that the IBD geographic pattern was not due to natural selection, we examined excess sharing across the genome for all IBD segments (SI Appendix) in European and North African IBD individuals. We detected a total of five regions with an excess of sharing when compared with the rest of the genome: in chromosome 6, coinciding with the HLA region, and at the tails of chromosomes 9, 17 and 19 (Fig. S4). We removed IBD segments within these regions and recalculated $W_{EA}$ between European populations and North Africa. The Pearson’s correlation of $W_{EA}$ results between the complete IBD dataset and the dataset without those regions that displayed extensive sharing was 0.99, which reinforces the robustness of our results. The maximum difference observed in $W_{EA}$ between Europe and North Africa was found in Hungary, with a 29% of difference between the two datasets.

A gradient of shared IBD segments is observed from southern to northern Europe (based on $W_{EA}$, Fig. 2, Table S2). This sharing is highest in the Iberian Peninsula for both North Africa and Sub-Saharan African IBD segments (with the exception of the Basques who show similar levels of sharing to other European populations). Additionally, we emphasize that the amount of IBD sharing detected between Sub-Saharan Africa and Europe is almost one order of magnitude lower than that between North Africa and Europe, and 30% of the IBD segments shared between Sub-Saharan Africa and Europe are also shared between North Africa and Europe. Interestingly, these segments represent only 2% of the bulk of IBD segments shared between North Africa and Europe, a proportion similar to that found in previous studies based only on Sub-Saharan populations (9). Considering that only 2% of the segments shared between North Africa and Europe have a Sub-Saharan origin, it is not likely that the gradients observed in Fig. 2A,B are driven primarily by the sub-Saharan segments. Finally, high correlation (0.83) exists on the values of $W_{EA}$ between Sub-Saharan Africa and Europe, and North Africa and Europe. Overall, these results support the hypothesis that Sub-Saharan gene flow detected in Europe entered with North African gene flow. We regressed the North African-European IBD metric ($W_{EA}$) on the sine of latitude to evaluate the strength of this gradient, and find a significant relationship across southern-to-northern Europe, $p=7.4\times10^{-8}$ (Fig. 3).

To pinpoint which specific North African regions exchanged migrants with Europe, we calculated $W_{EA}$ between a given European population and each of the 7 North African and Near Eastern populations (Fig. 4, Table S2). Southwestern European populations, and in particular the Canary Islands, show the highest levels of IBD sharing with northwestern African populations (i.e., the Maghreb: Morocco, Western Sahara, Algeria and Tunisia), whereas southeastern European populations share more IBD segments with Egypt and the Near East (Fig. S5). While inferred IBD sharing does not indicate directionality, the North African samples that have highest IBD sharing with Iberian populations also tend to have the lowest proportion of the European cluster in ADMIXTURE (Fig. 1), e.g. Saharawi, Tunisian Berbers and South Moroccans. For example, the Spanish share many IBD segments with the Tunisians and Western Saharans (Fig. 4), who present extremely minimal levels of European ancestry. This suggests that gene flow occurred from Africa to Europe rather than the other way around.
These results also rule out a model where observed sharing between Europe and North Africa is the result of recent gene flow from the Near East into both regions. We compared IBD between Qatari individuals (the best Near Eastern representatives genotyped with the Affymetrix platform currently available, Fig. S6) and both Europe and North Africa. As shown in Fig. 4 and Fig. S5, southwestern Europe has more IBD segments shared with the Maghreb than Qatar, whereas eastern Mediterranean populations share more segments IBD with the Near East than with western North Africa. On the other hand, northern European populations show only limited IBD sharing with both North Africa and the Near East (Fig. 2C, 4, S5). The southwest to northeast gradient of North African IBD sharing (Fig. 2A) and the distinct peak in sharing between Iberia and the Maghreb (Fig. 4) indicate that sharing in southwestern Europe is independent of gene flow from the Near East. It is possible that this sharp peak of North African IBD sharing in Iberia contributes to the apparent isolation of Iberian populations from other Europeans (43).

**ADMIXTURE vs. IBD estimated proportions:**

We next compared the frequency-based vs. haplotype sharing methods of estimating North African ancestry proportions. If gene flow has occurred in the distant past (i.e. more than 30 generations ago), ancestry detected by allele frequencies may exceed estimates of ancestry from recent long haplotypes IBD. Fig. S7, S8 show the correlations and proportions of sharing detected with ADMIXTURE for $k=4$ to $6$ compared with estimates from fastIBD. The correlation between IBD and allele-based estimates of North African ancestry is highest at $k=6$ ($R^2=0.80$). Allele estimates of joint Near Eastern and North African from $k=4$ substantially over-estimate IBD proportions. Concordance between IBD and $k=6$ allele-based estimates is particularly clear for southwestern European populations (Fig. S8).

However, for Central and North European populations, the $k=6$ cluster estimates of North African ancestry tend to be 0%, while haplotype sharing estimates average about 2% of the European genome. This difference is possibly due to either false positive matching with short IBD tracts (although all were greater than 1.5 cM) or remnants of Near Eastern ancestry that are also shared with North Africa. Short tract lengths on the order of a couple of centimorgans can persist in a population (or shared between populations) for thousands of years; for example, using the PopRes dataset, Ralph and Coop (43) show that a 2 cM tract shared between the French and British most likely traces back to a common ancestor more than 2,000 years ago.

**Implications of Gene Flow from North Africa to Europe**

**Time since admixture estimates:**

The variance in ancestry assignments for individuals within a population depends on the total ancestry proportions, the timing and duration of gene flow, population structure and/or assortative mating within the population, and errors in assignment. We used variance in ancestry proportions across individuals estimated with ADMIXTURE to infer effective admixture times, i.e. the times required to achieve the observed variance in the population given a single gene flow event in a randomly mating population (see model from (30)). Focusing on the North African component at $k=6$, we found that a migration event from North Africa to Europe would have occurred at least 6-10 generations ago (approximately 240-300 ya) in Spain, and at least 5-7 generations ago in...
France and Italy (Fig. 5). The pattern of North African ancestry at \( k = 7 \) remains very similar to the pattern at \( k = 6 \) with the estimate of admixture time decreasing 1 generation on average for Iberian populations (Fig. S13). Since population structure, continuous gene flow, assortative mating, and errors in assignments may considerably increase the variance (and thus reduce the effective migration time), we consider these time estimates to be lower bounds: under all the proposed variance-increasing scenarios, there must be a substantial proportion of migration that has occurred before the effective migration time, possibly much earlier. We additionally compare the estimate variance in ancestry from simulated populations to that predicted by a pulse model of migration. We found that the estimates were consistent with the actual number of generations since migration began, within confidence intervals obtained from bootstrapping over simulations (SI Appendix, Fig. S16, S17). Additionally, these estimates were robust to imperfect inference of the North African ancestry or source population when the pulse of gene flow occurred less than 15 generations ago.

**Disease Risks:**

We asked whether the migrations between North Africa and Europe affected the pattern of alleles associated with disease risk in these regions (44). By drawing on a database of GWAS risk alleles, we determined the cumulative risk for 134 diseases in each European and African population for which we had high-density SNP data (SI Appendix). We studied the deviations from random drift for all diseases with a False Discovery Rate (FDR) < 0.05. Pairwise \( q \)-values controlling for the FDR of all possible population comparisons within each disease (not across all diseases) were also calculated. The vast majority of disease alleles reflect expected patterns of neutral divergence (assessed with \( F_{ST} \)) among populations. Interestingly, we found that the multiple sclerosis (MS) risk calculated from 53 independent loci displayed a significant deviation from random drift for several North African populations. Maghrebi populations (e.g. Moroccans, Fig. S9) had a significantly elevated genetic risk for MS, while the Canary Islanders, the population with the highest inferred North African ancestry, had a significantly decreased risk for MS. We computed the cumulative genetic risk of each population using the 53 known SNPs associated with MS that intersect our data set. The Northern and Southern Moroccan populations have a cumulative risk allele frequency of 0.55 and 0.52, respectively. The Canary Islanders have a cumulative risk allele frequency of 0.44. This is beyond what is expected under genetic drift (FDR < 0.05). While MS prevalence is thought to increase along south-to-north latitudinal gradients in the northern hemisphere, prevalence data for North Africans are extremely limited (45). Our results suggest that North African Maghrebi have a greater genetic risk than expected under a neutral model, although presentation of MS could be attenuated by environmental variables such as UV exposure (46).

**Discussion:**

Using genome-wide SNP data from over 2,000 individuals, we characterize broad clinal patterns of recent gene flow between Europe and Africa that have a perceptible effect on genetic diversity of European populations. We have shown that recent North African ancestry is highest in southwestern Europe and decreases in northern latitudes,
with a sharp difference between the Iberian Peninsula and France, where Basques are less influenced by North Africa (as suggested in (47)). Our estimates of shared ancestry are much higher than previously reported (up to 20% of the European individuals’ genomes). This increase in inferred African ancestry in Europe is due to our inclusion of 7 North African, rather than Sub-Saharan African populations. Specifically, elevated shared African ancestry in Iberia and the Canary Islands can be traced to populations in the North African Maghreb like Moroccans, Western Saharan and the Tunisian Berbers. Our results, based on both allele-frequencies and long shared haplotypes, support the hypothesis that recent migrations from North Africa contributed substantially to the higher genetic diversity in southwestern Europe. Previous Y-chromosome data have highlighted examples of male-biased gene flow from Africa to Europe, such as the eastern African slave ancestry in Yorkshire, England (48) and the legacy of Moors in Iberia (49). Here we show that gene flow from Africa to Europe is not merely reflected on the Y-chromosome but corresponds to a much broader effect.

Alternative models of gene flow: Migration(s) from the Near East likely have had an effect on genetic diversity between southern and northern Europe (discussed below), but do not appear to explain the gradients of African ancestry in Europe. A model of gene flow from the Near East into both Europe and North Africa, such as a strong demic wave during the Neolithic, could result in shared haplotypes between Europe and North Africa. However, we observe haplotype sharing between Europe and the Near East follows a southeast to southwest gradient, while sharing between Europe and the Maghreb follows the opposite pattern (Fig. 2); this suggests that gene flow from the Near East cannot account for the sharing with North Africa.

We do detect low levels of IBD and allele sharing between the Near East and the majority of the European continent. Both IBD and allele sharing with the Near East appear elevated in southeastern Europe (e.g. Italy, Yugoslavia, Cyprus). It is possible that these patterns reflect more ancient migrations, perhaps dating back to the Neolithic, which resulted in a low level of short Near Eastern haplotypes across much of Europe. This hypothesis is further supported by results of the time since admixture estimate based variances in ancestry proportions (Fig. 5), which suggest that Near Eastern ancestry is older than the North African one and therefore they did not enter Europe on the same migration wave.

Another possible hypothesis to explain the increased diversity in southern Europe could have been that an influx of Jewish ancestry had a heterogeneous effect on genetic diversity in Europe. However, in most European populations here, virtually no Jewish ancestry was detected. On average, 1% of Jewish ancestry is found in Tuscan HapMap population and Italian Swiss, as well as Greeks and Cypriots. This may reflect the higher sharing with Near Eastern populations in the Italian peninsula and southeastern Europe (Fig. 2C) or low levels of gene flow with the early Italian Jewish communities (28). Estimates from the IBD analysis are in agreement with ADMIXTURE estimates that the amount of sharing between these populations is extremely low (Table S2). Specifically, results of IBD sharing between southwestern Europe and North Africa are two orders of magnitude greater than those found between the same region and Jews, the average W_{EA} for southwestern Europe and North Africa is 203, while for southwestern Europe and European Jews it is 1.3.
Disease Risk Implications:
The observation that the majority of disease risk alleles in this study follow an expected pattern of neutral drift among populations is consistent with the interpretation that these common alleles are not strongly affected by natural selection. We note that alleles identified in GWASs of individuals of largely northern European descent have limited portability to neighboring populations because the tagged GWAS SNPs may no longer be in linkage disequilibrium with the causative variant. Thus, estimates of genetic risk for these diseases in North Africans are likely inaccurate because North African specific risk SNPs are missing. With these caveats, we note that one disease, multiple sclerosis does not conform to a pattern of neutral genetic drift and this raises the hypothesis that natural selection affects the frequency of these risk variants that may also be linked to phenotypes other than MS. Our results show an increased genetic risk for multiple sclerosis in North African populations. West Saharans and North Moroccans carry higher frequencies of MS alleles that deviate from neutral expectations of divergence among European and African populations. Based on our model, we would predict individuals with high North African ancestry living in Europe to have a higher genetic risk for MS (see supporting evidence for North African immigrants in France in (50)). However, the Canary Islands, while displaying the highest amount of North African ancestry, have the lowest predicted genetic risk for MS. The complexity of these results serves to emphasize the importance of conducting disease associations in many diverse populations (51). The significant gene flow from North Africa into southern Europe will result in a miscalculation of genetic disease risk in certain European populations, if North African specific risk variants are not taken into account.
Materials and Methods:

Data: Recently published and new single nucleotide polymorphism (SNP) data were used to build a database of 43 populations and 2,099 individuals. The database includes 7 North African populations (25), together with data from 27 European populations (25, 27, 52), 2 European Jewish populations (28), 1 Near Eastern population (29) and HapMap3 Sub-Saharan African populations (Table S1). Additionally, new data for 3 Spanish populations: Galician (NW Spain), Andalusian (S Spain), and the Canary Islands, was included in the database. Informed consent was obtained from all newly collected Spanish populations. Samples were genotyped on the Affymetrix 6.0 chip, and quality control filtered for missing loci and close relatives. Data from these new populations can be found at: bhusers.upf.edu/dcomas/

ADMIXTURE analysis: An unsupervised clustering algorithm ADMIIXTURE 1.21 (33) was used to determine allele-based sharing in a dataset of 243,000 markers formed by a total of 41 populations (29 Europeans, 2 Jewish Europeans, 1 Near Eastern, the Qatari, 7 North Africans and 2 Sub-Saharan Africans). For the sake of equal representation, a random subset of 15 individuals was chosen for any population having a much larger sample size. Ten ancestral clusters (k=2 through 10) in total were tested successively, running 10 iterations for each ancestral cluster (Fig. S1) and calculating cross validation errors for every run (Fig. S2). Moreover, for k=4 through 6,200 bootstraps were performed by resampling subsets of each chromosome, so that standard errors for each ancestral cluster estimate could be obtained (33).

RFMix analysis: We used a second program “RFMix” (36) to estimate admixture proportions using defined source populations. After phasing the European and North African populations with BEAGLE, the default settings in RFMix were employed to minimize switch errors. Three possible source populations were used to estimate European ancestries: a northern European population (Germans), North African (Tunisians), and Near Eastern (Qatari). A forward-backward algorithm was used to generate posterior ancestry probabilities at each SNP. Setting a posterior confidence
threshold of 99%, we determined the mean proportion of SNPs with max-marginal probabilities above this threshold for each ancestry in each population.

**Modeling migration timing**: According to Gravel *et al.* (30) the relationship between variance in ancestry \( \text{Var}(X^p) \), time since admixture \( T \) and migration proportions \( m \) is simply:

\[
\text{Var}(X^p) = \frac{m(1-m)}{2^{T-1}} + \frac{2m(1-m)(1-1/2N)^{T-1}}{2n + 2(T-2)L}
\]

where \( L \) is the total length of the genome (in Morgans), \( n \) is the number of chromosomes [22], and \( N \) is the population size. Here we used \( n=22, L=3,500 \text{ cM} \) and an infinite population size (as drift is a negligible force during very recent migrations).

**IBD detection**: The analysis of IBD sharing was conducted using all the populations in the dataset (Table S1) with the exclusion of the European Jewish populations. We note that in the *ADMIXTURE* analysis at \( K=3 \) there is shared ancestry between Europeans and Jewish populations, however, this could represent either shared ancestral variation or gene flow. Higher levels of \( K>3 \) showed very little recent Jewish ancestry in European populations and North African populations show negligible ancestry from North African Jews (35). The removal of Jewish populations from the dataset increased the number of common markers from 243,000 to 274,000 and to a total of 41 populations. A preliminary test of IBD sharing was calculated with both GERMLINE (53) and fastIBD (40). Results showed fastIBD as the more accurate method to detect IBD in our dataset (SI Appendix), so that further analyses were done using this algorithm. We ran 10 iterations for each of the autosomal chromosomes to account for phasing uncertainty and we used a threshold of \( 10^{-10} \) for IBD detection, so that segments with a smaller score were included. This threshold showed better results than using \( 10^{-8} \) or \( 10^{-6} \) scores, based on a comparison of the segments detected with a High Density Dataset (HDD) and with a Low Density Dataset (LDD) (data not shown). After this, we excluded all those segments shorter than 1.5 cM in order to avoid an excess of false positives.
Correction for sample size: In order to compare between the different statistics calculated from the IBD results, we correct for sample size, given that in European populations there are differences in sample size of two orders of magnitude. We follow Atzmon et al. (28) calculation of IBD sharing metrics. Suppose we want to calculate a parameter related to the IBD between PopA and PopB. Our environment will be a list of all the individuals belonging to PopA and PopB. The correction factor is the total number of possible pairs where one individual is from PopA and the other is from PopB, which will depend on the population sizes, $n$ and $m$, respectively. This is:

$$W_{ab} = \frac{\sum \sum W_{ab}}{nm}$$

Standard deviation from $W_{ab}$ statistic was obtained on the basis of the standard deviation in IBD sharing between PopA and PopB. Estimates were scaled by 100 for ease of presentation.

Acknowledgements: We are grateful to Oscar Lao and Graham Coop for helpful suggestions. We extend our gratitude to the North African and Spanish participants for their generous contributions of DNA. We would like to thank the Banco Nacional de ADN (www.bancoadn.org/) for providing the Galician and Andalusian samples. BMH and CDB were supported by NIH grant 3R01HG003229. LRB and DC were supported by MCINN grant CGL2010-14944/BOS and Generalitat de Catalunya grant 2009SGR1101. CF was supported by by ISCIII grant PI11/00623. The Spanish National Institute for Bioinformatics (www.inab.org) supported this project, and we are grateful to Txema Heredia for IT help.
Figure Legends:

**Figure 1 Legend:** Allele-based estimates of ancestry in Europe, North Africa, Sub-Saharan Africa, Jews and the Near East. Unsupervised ADMIXTURE results for $k=3.6$. Cross-validation indicated $k=4$ as the best fit, but higher density datasets (25) and higher values of $k$ continue to identify population-specific ancestries (Fig. S2); we therefore conservatively focused on $k=3.6$ ancestral populations.

**Figure 2 Legend:** Haplotype-based estimates of genetic sharing between Europe and Africa shows that the highest sharing is with the Iberian Peninsula. Genetic sharing between geographic regions is represented as a density map of $W_{EA}$ estimates for 30 European populations where haplotypes are IBD with (a) Sub-Saharan Africa, (b) North Africa and (c) the Near East. The Canary Islands are shown in the bottom left.

**Figure 3 Legend:** Significance of latitudinal gradient of IBD sharing within Europe. In order to determine if there was a significant relationship between latitude and mean IBD count ($W_{EA}$) within Europe, we regressed $W_{EA}$ on $\text{log}(\sin(\text{latitude}))$. The sine of the latitude was used to obtain distance-appropriate vertical values; then we log-transformed these value to obtain the expected decay of allele sharing in 2-dimensional habitats (54). The p-value of the regression for IBD shared between North Africa and Europe is $7.4 \times 10^{-8}$ (the p-value for sharing between Europe and Sub-Saharan Africa is $8.7 \times 10^{-9}$, data not shown).

**Figure 4 Legend:** Population-specific estimates of haplotype sharing (cM) between North Africa and Europe. Estimates of $W_{EA}$ between each European population (labeled on the X-axis) and each of the 7 North African populations and the Qatari are represented by colors and symbols. Estimates of $W_{EA}$ were scaled by 100 for ease of presentation. A substantial increase in haplotype sharing is detected between southwestern European populations and Maghrebi populations (i.e. Morocco, Western Sahara and Tunisia) in comparison to the remainder of the European continent. The excess of sharing between the Near East and southern central and Eastern Europe is also noteworthy.
**Figure 5 Legend:** *Variance in ancestry proportions within populations depends on the overall ancestry proportions in the population and the time of gene flow.* (a) Using the proportion of Near Eastern | North African ancestry inferred at $k=4$ with *ADMIXTURE*, we estimated the variance in ancestry within each of 11 European populations, indicated here by abbreviations. The grey lines show the expected relationship between ancestry proportions (X-axis) and variances (left Y-axis), under a single pulse model occurring at generation $g$ (right Y-Axis). Departures from single-pulse models tend to increase the variance in ancestry and so the corresponding effective times should be thought of as lower bounds: significant migration must have occurred before the effective times (see text). (b) Estimating the effective time of migration based on variance in North African ancestry proportions inferred under the $k=5$ model (Fig. 1). (c) Estimating the effective time of migration based on variance in North African ancestry proportions inferred under the $k=5$ model.
References:


43. Ralph P & Coop G (arXiv:1207.3815 [q-bio.PE]).


