SUPPLEMENTARY INFORMATION

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

Laura R. Botigué[1*], Brenna M. Henn[2*], Simon Gravel[2], Brian K. Maples[2], Christopher R. Gignoux[3], Erik Corona[4,5], Gil Atzmon[6,7], Edward Burns[6], Harry Ostrer[7,8], Carlos Flores[9,10], Jaume Bertranpetit[1], David Comas[1¤], Carlos D. Bustamante[2¤]

INDEX

Supplementary Methods

RFMix: RFMix was run in order to estimate ancestral proportions in the mode that accounts for possible switch errors using default settings and population phased data. Also, the forward-backward algorithm was used to generate posterior ancestry probabilities at each SNP. Setting a confidence threshold of 99%, we determined the mean proportion of SNPs with max-marginal probabilities above this threshold for each ancestry in each population.

ADMIXTURE simulations: To simulate admixed chromosomes, we followed a two-step process. In the first step, we created ancestry tracts in a diploid Wright-Fisher population of 1000 individuals over 15 generations, and extracted 50 individuals at generations G=5, 10, and 15. As a result, ancestry assignments for each SNP in the paternal and maternal haplotypes of each individual were obtained.
In the second step, we constructed the simulated individuals in the following way: We performed 10 iterations where we randomly chose 5 CEU and 5 Tunisians as reference ancestors to create 5 simulated individuals using the ancestry tracts of 5 sampled individuals of the simulated admixed population built in the former step. The reference ancestors were removed from the population panel used to run ADMIXTURE, to avoid a false relatedness with the simulated individuals. Finally, for each iteration, the population panel was merged with the 5 simulated individuals and the resulting dataset was used to run ADMIXTURE.

IBD detection: An initial test of IBD sharing was calculated with both GERMLINE (1) and fastIBD (2). For GERMLINE default settings were used, except for the following flags: *-min_m* and *-err_het* were set to 1, *-bits* were set to 120, and the *-w_extend* was activated. For fastIBD algorithm, a fastIBD-score threshold of 10^{-10} was chosen and final results were the combination of 10 runs, as recommended in Browning *et al*. (2). As an initial test, we attempted to identify segments found in chromosome one, of at least 1.5 cM in length, and shared between Europe and North Africa. A total of 20,080 segments were detected between Europe and North Africa when using GERMLINE, whereas only 6,208 were found by fastIBD. Table S4 shows the mean number of IBD segments detected for each European population, corrected by sample size (see *Supplementary Results* below). Differences between methods in the amount of sharing were maintained at a population level. Overall, GERMLINE results did not seem to reflect a geographic pattern of sharing in European populations, but rather similar sharing among populations, except for some outliers with an increased (Greece, Finland) or decreased (Romania) amount of IBD sharing with North Africa. fastIBD results showed that Southwestern Europeans had the highest amounts of sharing with North Africa. GERMLINE has higher power and a low false positive discovery rate when detecting segments of a minimum of 4 cM length (2). Thus, GERMLINE is more suitable for the analysis of IBD between closely related individuals. On the other hand, fastIBD has high power and low false discovery rate when detecting segments of at least 1.5 cM length, which is a more suitable length for our time depth. The increased number of short IBD segments detected with GERMLINE, but not found with fastIBD, is likely due to its higher false positive rate; we thus conducted further analysis using fastIBD.

Sensitivity of IBD Metrics: We 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 with a low-density dataset (LDD) of 280,462 SNPs from our primary analysis. We calculated the proportion of IBD segments detected in LDD that are also present in HDD, and found that when we considered only North African vs. European IBD segments the proportion was 72%, whereas when we considered only segments found within Europe the proportion was 80% (Fig. S15). This latter value coincides with the power of BEAGLE using the same fastIBD threshold (10^{-10}) in European populations (2). The HDD detects 4x more segments than the LDD (63,368 and 15,939 segments, respectively). This increase in the number of segments is not surprising considering the power of additional markers to detect short, shared segments. Moreover, the segment length distribution of the HDD dataset is exponentially shaped (in contrast to the LDD) as expected due to the decay in length under a Poisson process of recombination (Fig. S16). Average segment lengths are also slightly different between the datasets, 2.63 cM in the HDD and 3.88 cM in the LDD, suggestive that higher density markers better capture the edges of a given shared segment.

Detecting IBD peaks: An excess of IBD sharing in a given region may be caused by a share effect of positive selection in the two populations (3). To confirm that the IBD geographic pattern was not due to the effects of adaptation, we examined excess sharing across the genome for all IBD segments in European and North African IBD individuals. We detected a total of five regions with an excess of sharing 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. S6). 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.

Risk allele frequencies: We asked whether the migrations between North Africa and Europe affected the pattern of alleles associated with disease risk in these regions. Using a database developed in (4) after manual curation from published literature, SNPs associated with a disease in a genome wide association study (GWAS) and having a p-value below $1x10^{-6}$ with a reported risk allele were included in this analysis. Candidate gene studies were not included due to the large p-values that are reported and the resulting skepticism they cause. Only single SNP GWAS hits were included (as opposed to haplotype blocks associated with a disease). In cases where different disease risk alleles were reported for the same disease in different studies, the risk allele in the study with the largest sample size (disease + non-disease individuals) was used. Since many GWAS SNPs are in linkage disequilibrium (LD) with the actual causal SNP, we filtered SNPs with LD r^2 value greater than 0.2 to insure only one SNP per associated region was used. SNPs with the largest odds ratio were retained during local filtering as they are more likely to reflect the risk associated with the actual causal SNP. When the odds ratio was not reported, retention of SNPs with the largest sample size in the study were prioritized. Cumulative risk allele frequency results for each population and 134 diseases are plotted online at geneworld.stanford.edu/africa_hapmap. The cumulative risk allele

frequency is number of risk alleles present in each population across all SNPs associated with the disease divided by the total number of alleles.

Supplementary Results and Discussion

ADMIXTURE: Cross validation errors were lowest at k=4 (Fig. S2), where ancestry assignment differentiated the European populations, European Jews, Sub-Saharan Africans, and a cluster containing Near Eastern and North African populations. Previous work with a denser SNP dataset has shown (5) that higher k values consistently pull out population-specific clusters (e.g. Jews, Tunisians, Basque), therefore we present results for ancestral populations greater than the cross-validation minimum of k=4. The presence of a cluster found almost exclusively in Jewish populations is not surprising when considering the high extent of their shared IBD segments (6). It is interesting to note that this ancestry is not represented in European individuals. Jewish ancestry appears to be less than 2% or absent in most populations, with the exception of one Swiss Italian. At k=6, a component corresponding largely to North Africa appears, except for the Tunisians who are ~100% assigned to their own component, likely due to strong endogamy (5). The differentiation of Tunisians reduces the genetic similarity between Near Easterners and North Africans when comparing their ancestry assignments at k=5 and k=6. The Basques have the lowest proportion, only 4% is assigned North African ancestry. However, we detect that for all iterations Basques are represented by a single ancestry at values of k=6:10 (Fig. S1), in agreement with previous studies (7, 8). Moreover, it is interesting to notice that this Basque ancestry ranges between 50 - 11% of the genome in the remaining Iberian Peninsula populations as well as French and Italian ones, suggesting the existence of a Southern European component.

ADMIXTURE simulations: We note that correlation at the population level between ancestry proportions estimated by ADMIXTURE and simulated ancestry proportions is high, and discrepancies at the individual level are at the order of ±2%. Using the inferred ancestry proportions in these simulated individuals, we compared the estimated variance in ancestry to that predicted by a pulse model of migration. We found that the estimates from G=5 and G=10 were consistent with the actual number of generations, within the confidence intervals obtained by bootstrap. By contrast, the method underestimates the age of an admixture starting 15 generations ago. This gives us confidence that our method would have been able to detect traces of recent (<10 generations) admixture (Fig. S12, S13).

Length of IBD Segments: We calculated a second statistic "L_{EA}", the average length of the segments shared IBD between a pair of individuals, one from European population and the other from North Africa | Sub-Saharan Africa. Normalization was based on the possible number of pairwise comparisons between both populations. L_{EA} reflects the time since gene flow occurred in contrast to W_{EA} (6). Interestingly, L_{EA} shows an opposite pattern, northern and central European populations having higher values than southern ones (Fig. S17). This suggests that gene flow between southern Europe and North Africa is older than that in other regions in Europe, where longer (recent) segments are found. 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. This suggests that gene flow occurred from Africa to Europe rather than the other way around.

Jewish ancestry in Europe: Another possible hypothesis to explain the increased diversity in southern Europe is 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 (6). Estimates from the IBD analysis are in agreement with ADMIXTURE estimates that the amount of sharing between these populations is extremely low (SI Appendix, Table S3). 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 WEA for southern Europe and North Africa is 203, while for southwestern Europe and European Jews is 1.3.

Supplementary Tables

Table S1 Description of the dataset.

Population	PopID	Size	Region	Ref.
Morocco North	MorN	18	North Africa	3
Morocco South	MorS	16	North Africa	3
Occidental Sahara	Sah	18	North Africa	3
Algeria	Alg	19	North Africa	3
Tunisia Berbers	Tun	18	North Africa	3
Libya	Lib	17	North Africa	3
Egypt	Egy	19	North Africa	3
Canary Islands	Can	17	SW Europe	Present
Spain Andalucia	And	17	SW Europe	Present
Spain Galicia	Gal	17	SW Europe	Present
Spain Basques	Bas	20	SW Europe	3
Spain General	Spa	48	SW Europe	6
Portugal	Por	117	SW Europe	6
France	Fra	89	S Europe	6
Italy	Ita	108	S Europe	6
Italy Tuscan	TSI	88	S Europe	9
Yugoslavia	Yug	8	SE Europe	6
Greece	Gre	2	SE Europe	6
Cyprus	Cyp	1	SE Europe	6
Belgium	Bel	42	C Europe	6
Netherlands	Net	16	C Europe	6
Germany	Ger	69	C Europe	6
Austria	Aus	11	C Europe	6
Switzerland German	SzG	84	C Europe	6
Switzerland Italian	SzI	13	C Europe	6
Switzerland French	SzF	754	C Europe	6
CEU	CEU	88	NW Europe	9
Ireland	Ire	4	NW Europe	6
Scotland	Sco	2	NW Europe	6
United Kingdom	UK	24	NW Europe	6
Hungary	Hun	1	NE Europe	6
Romania	Rom	1	NE Europe	6
Poland	Pol	18	NE Europe	6
Russia	Rus	6	NE Europe	6
Finland	Fin	1	NE Europe	6
Sweden	Swe	10	NE Europe	6
Norway	Nor	2	NE Europe	6
Italian Jews	ItaJ	15	Europe	7
Ashkenazi Jews	AshJ	15	Europe	7
Qatari	Qat	20	Near East	8
Nigeria Yoruba	YRI	100	W Sub-Saharan Africa	9
Kenya Luhya	LWK	90	E Sub-Saharan Africa	9
Kenya Maasai	MKK	56	E Sub-Saharan Africa	9

Table S2: *Estimates of European, North African and Near Eastern Ancestry in Europeans using RFMix*

ADMIXED POPULATION	SOURCE POPULATIONS			
	German	**Saharawi**	**Qatari**	**UnCalled**
Spain Canary Islands	0.44	0.14	0.13	0.29
Portugal	0.48	0.10	0.14	0.29
Spain Galicia	0.48	0.09	0.14	0.28
Spain Andalucia	0.48	0.09	0.14	0.28
Spain Central	0.49	0.09	0.14	0.28
Italy	0.46	0.07	0.20	0.28
Spain Basque	0.51	0.07	0.13	0.29
Cyprus	0.40	0.07	0.26	0.27
Swiss Italian	0.50	0.06	0.17	0.28
Tuscan	0.47	0.06	0.19	0.28
Hungary	0.54	0.06	0.12	0.28
Greece	0.45	0.06	0.22	0.27
France	0.53	0.05	0.14	0.28
Swiss French	0.53	0.05	0.14	0.28
Swiss German	0.54	0.05	0.14	0.28
Yugoslavia	0.52	0.05	0.16	0.27
Belgium	0.55	0.05	0.13	0.27
Austria	0.55	0.05	0.13	0.27
UK	0.57	0.04	0.12	0.27
CEU	0.57	0.04	0.12	0.27
Russia	0.56	0.04	0.12	0.28
Romania	0.53	0.04	0.18	0.25
Netherlands	0.58	0.04	0.12	0.27
Norway	0.57	0.04	0.10	0.29
Ireland	0.57	0.04	0.11	0.27
Poland	0.58	0.04	0.11	0.27
Sweden	0.58	0.04	0.11	0.27
Finland	0.57	0.04	0.11	0.29
Scotland	0.57	0.04	0.11	0.28

Table S3 *W_{EA} statistic and standard deviation between European populations and North Africa, the Near East and sub-Saharan Africa.*

	NAfrica	NEast	SubSah	MorN	MorS	OccSah	Alg	Tun	Lib	Egy	Jew
CAN	349 ± 54	152 ± 70	21 ± 14	388 ± 77	401 ± 85	465 ± 168	199 ± 65	424 ± 86	254 ± 91	104 ± 53	1.51 ± 1
POR	227 ± 43	110 ± 51	9 ± 24	283 ± 79	246 ± 82	252 ± 86	144 ± 58	266 ± 115	166 ± 60	75 ± 44	1.23 ± 1
GAL	204 ± 34	110 ± 47	4 ± 2	254 ± 97	241 ± 74	225 ± 70	106 ± 33	236 ± 101	158 ± 45	86 ± 52	0.40 ± NA
AND	186 ± 33	107 ± 38	7 ± 4	236 ± 82	214 ± 86	198 ± 64	97 ± 33	251 ± 95	121 ± 45	78 ± 40	2.72 ± 2
SPA	192 ± 41	108 ± 52	5 ± 3	227 ± 84	196 ± 70	204 ± 72	129 ± 54	234 ± 91	148 ± 69	76 ± 37	1.32 ± 1
BAS	62 ± 20	59 ± 33	1 ± 1	86 ± 41	45 ± 30	62 ± 40	43 ± 23	57 ± 41	51 ± 25	42 ± 25	0.69 ± 0
FRA	74 ± 21	84 ± 46	1 ± 1	88 ± 47	61 ± 37	68 ± 46	48 ± 33	78 ± 43	63 ± 35	53 ± 31	1.08 ± 1
ITA	98 ± 31	170 ± 76	4 ± 3	103 ± 60	76 ± 45	86 ± 48	69 ± 39	93 ± 56	99 ± 55	85 ± 38	0.91 ± 1
TSI	71 ± 18	120 ± 58	2 ± 1	78 ± 43	52 ± 33	69 ± 35	42 ± 30	63 ± 41	68 ± 38	74 ± 38	1.19 ± 1
GRE	85 ± 11	250 ± 167	6 ± 4	74 ± 75	65 ± 28	46 ± 22	61 ± 0	138 ± 37	44 ± 25	115 ± 9	4.59 ± 2
YUG	55 ± 10	104 ± 39	1 ± 0	81 ± 49	41 ± 33	19 ± 11	36 ± 15	39 ± 19	71 ± 32	57 ± 34	2.58 ± 1
CYP	*103*	*242*	*6*	*94*	*116*	*97*	*51*	*78*	*57 ± NA*	*154 ± NA*	0.00 ± NA
BEL	59 ± 17	74 ± 38	1 ± 1	78 ± 37	48 ± 34	44 ± 32	32 ± 16	64 ± 40	49 ± 36	53 ± 30	0.97 ± 0
NTL	47 ± 14	65 ± 39	1 ± 1	74 ± 37	29 ± 14	32 ± 18	34 ± 26	45 ± 26	35 ± 21	33 ± 29	0.35 ± NA
GER	55 ± 17	84 ± 47	1 ± 1	80 ± 33	45 ± 28	44 ± 34	30 ± 19	50 ± 35	51 ± 33	43 ± 26	2.32 ± 2
OST	57 ± 16	96 ± 39	1 ± 1	76 ± 32	42 ± 24	40 ± 35	43 ± 18	36 ± 21	49 ± 18	54 ± 23	0.00 ± NA
SWZ_GE	57 ± 15	87 ± 49	1 ± 1	74 ± 41	42 ± 27	40 ± 31	40 ± 26	61 ± 44	53 ± 30	47 ± 26	0.75 ± 0
SWZ_IT	79 ± 39	142 ± 88	2 ± 2	74 ± 46	67 ± 37	75 ± 48	54 ± 35	67 ± 46	77 ± 73	77 ± 41	6.57 ± 10
SWZ_FR	63 ± 18	93 ± 49	1 ± 1	79 ± 41	50 ± 32	53 ± 34	42 ± 27	62 ± 38	59 ± 36	53 ± 33	1.11 ± 1
CEU	53 ± 16	70 ± 45	2 ± 2	78 ± 48	41 ± 23	40 ± 26	34 ± 23	49 ± 37	45 ± 32	48 ± 33	1.08 ± 1
IRE	48 ± 18	56 ± 37	1 ± 0	70 ± 52	37 ± 26	12 ± 5	15 ± 8	54 ± 31	43 ± 23	68 ± 36	0.00 ± NA
SCO	41 ± 2	55 ± 15	0 ± NA	63 ± 76	9	19	13 ± 3	29 ± 29	53 ± 39	72 ± 14	3.29 ± 1
UK	51 ± 12	71 ± 45	1 ± 1	69 ± 27	38 ± 26	45 ± 26	36 ± 25	42 ± 24	43 ± 19	42 ± 22	1.95 ± 2
HUN	*78*	*58*	*2*	*97*	*69*	*27*	*70*	*60*	*91*	*128*	*0.00 ± NA*
ROM	*61*	*98*	*0*	*102*	*0*	*27*	*67*	*63*	*29*	*18*	*0.00 ± NA*
POL	55 ± 17	84 ± 46	1 ± 1	69 ± 36	47 ± 32	32 ± 30	42 ± 22	53 ± 32	55 ± 27	49 ± 28	4.13 ± 3
RUS	51 ± 18	69 ± 42	2 ± 2	49 ± 32	46 ± 26	47 ± 21	34 ± 23	34 ± 34	64 ± 51	38 ± 18	3.22 ± 2
FIN	*20*	*78*	*0*	*14*	*0*	*13*	*33*	*27*	*15*	*19*	*0.00 ± NA*
SWE	46 ± 20	75 ± 58	2 ± 2	64 ± 41	37 ± 34	24 ± 15	24 ± 12	51 ± 38	34 ± 21	47 ± 29	0.73 ± 0
NOR	40 ± 6	49 ± 30	0	69 ± 2	34 ± 13	40 ± 40	14 ± 6	41 ± 16	6	45 ± 43	1.58 ± NA

Populations with only one representative are shown in italics. Note that standard deviation is not shown when only one segment is detected.

Table S4 *Comparison of Germline and fastIBD at 1.5cM threshold*

PopID	Germline[1]	fastIBD[1]	PopID	Germline[1]	fastIBD[1]
AND	9.22	4.94	**OST**	7.78	1.31
BAS	8.00	2.40	**SWZ_DEU**	9.86	1.58
CAN	10.68	10.45	**SWZ_IT**	8.92	1.97
GAL	9.18	6.49	**SWZ_FR**	9.38	1.83
PTG	10.69	6.80	**CEU**	8.36	1.46
SPA	9.97	6.07	**IRE**	8.00	1.00
FRA	9.078	2.05	**SCO**	7.60	4.00
ITA	9.06	3.04	**UK**	9.10	1.77
TSI	9.74	2.30	**HUN**	6.40	0.80
YUG	9.30	0.60	**ROM**	4.80	1.60
GRE	14.40	2.00	**PLK**	8.76	1.51
CYP	10.40	5.60	**RUS**	8.99	2.13
BEL	10.17	2.13	**FIN**	15.2	0
NTL	8.40	1.40	**SWE**	8.32	1.52
DEU	9.27	1.37	**NOR**	6.80	1.20

[1] Number of segments longer than 1.5 cM detected to be shared between a given European population and North Africa corrected by sample size (see *Methods*).

Supplementary Figures

Figure S1

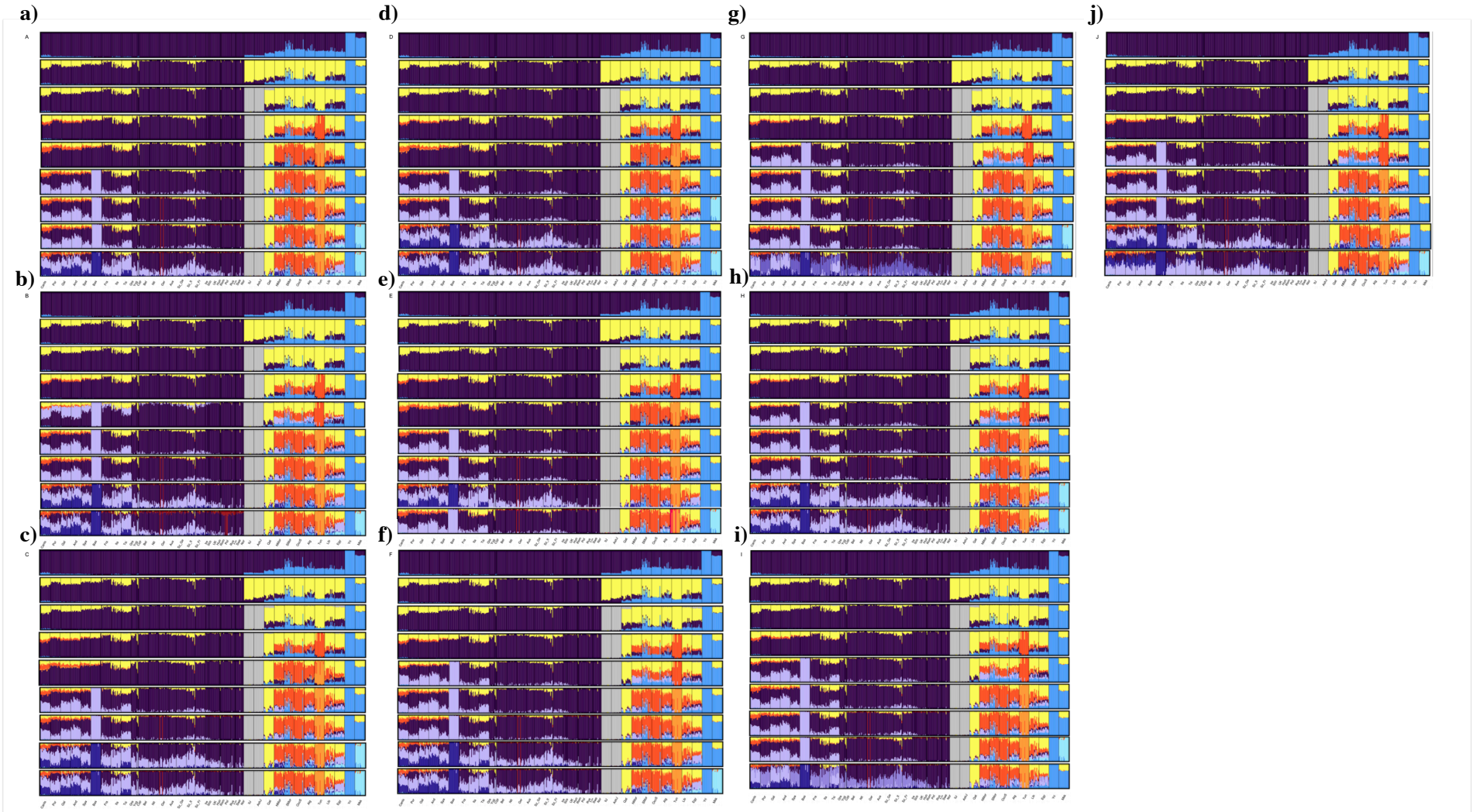

Figure S1: *ADMIXTURE iterations from k= 2 through 10*. Individuals are represented as vertical lines, and each k ancestral genetic cluster is represented by a color. No perceptible variation in the assignment of ancestries is detected from k= 2 to 5 across all iterations. At k = 6 two clusters arise depending on the iteration, either a new genetic cluster mainly assigned to Basques and Southern Europeans or a genetic cluster differentiating Tunisian populations from the other North African populations. Lowest CV errors at k = 6 are found when the Tunisian cluster appears. For k > 7, more variation in the ancestry assignments is observed across iterations.

Figure S2

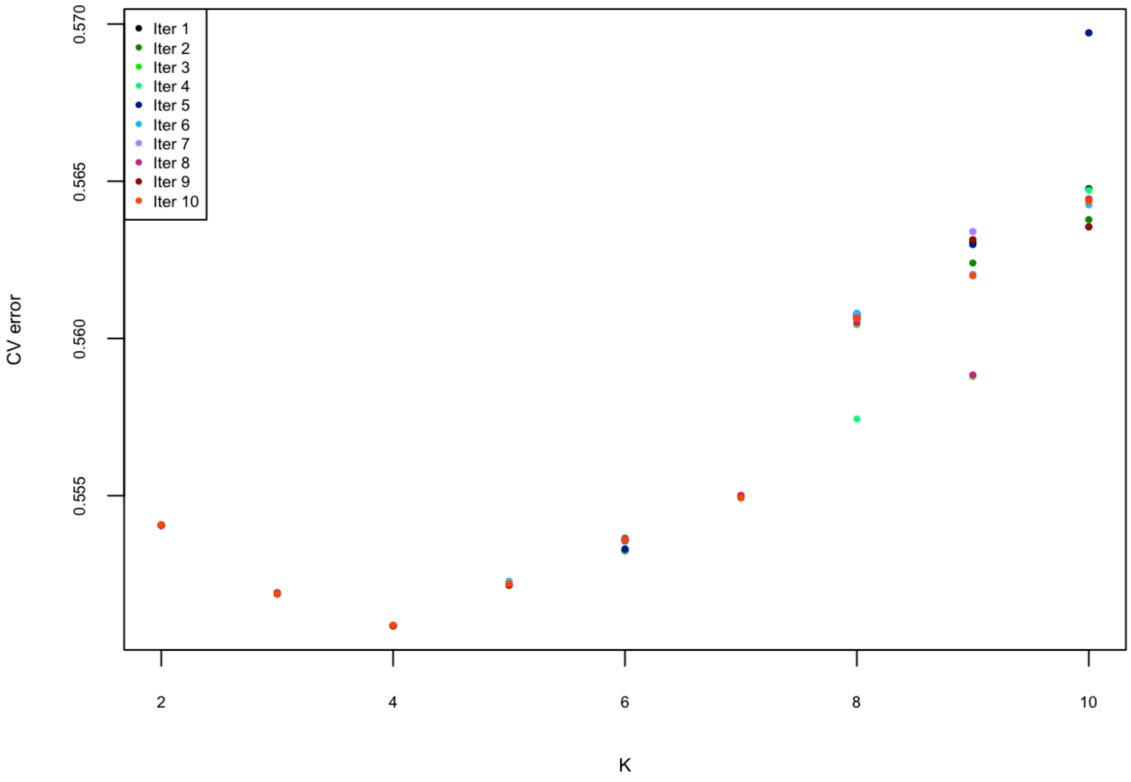

Figure S2: *10-fold cross validation error for ADMIXTURE assignment results*. On the x- axis results for each k ancestral population are represented, and every dot corresponds to a different iteration. Cross validation errors are lowest at k = 4, where ancestral cluster assignment allows the differentiation between the Sub-Saharan populations, the Europeans, Jewish populations and a group formed by Near Eastern and North African populations. However, the cross validation error is quite conservative, and we retain that real genetic structure exists between North Africa and Near Eastern populations, as well as between Tunisian and the rest of North Africans (given its known endogamy as shown in Henn et al. (4)). Thus, ancestry assignment for k= 4 to 6 was considered for discussion.

Figure S3

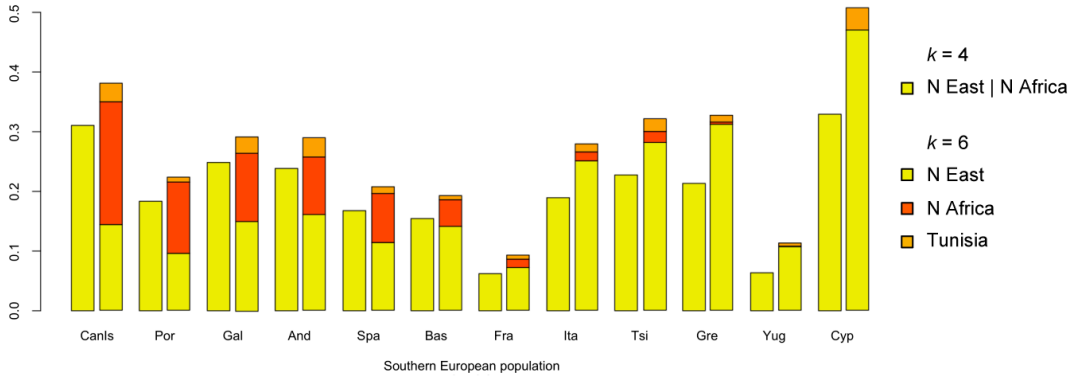

Figure S3: *ADMIXTURE estimated average proportions*. Near Eastern/North African ancestry is shown in yellow (k=4) and Near Eastern, North African and Tunisians ancestries in yellow, red and orange (k=6) in European populations. Bootstrap resampling resulted in an average individual standard error of ± 1.6% in assigning North African ancestry.

Figure S4

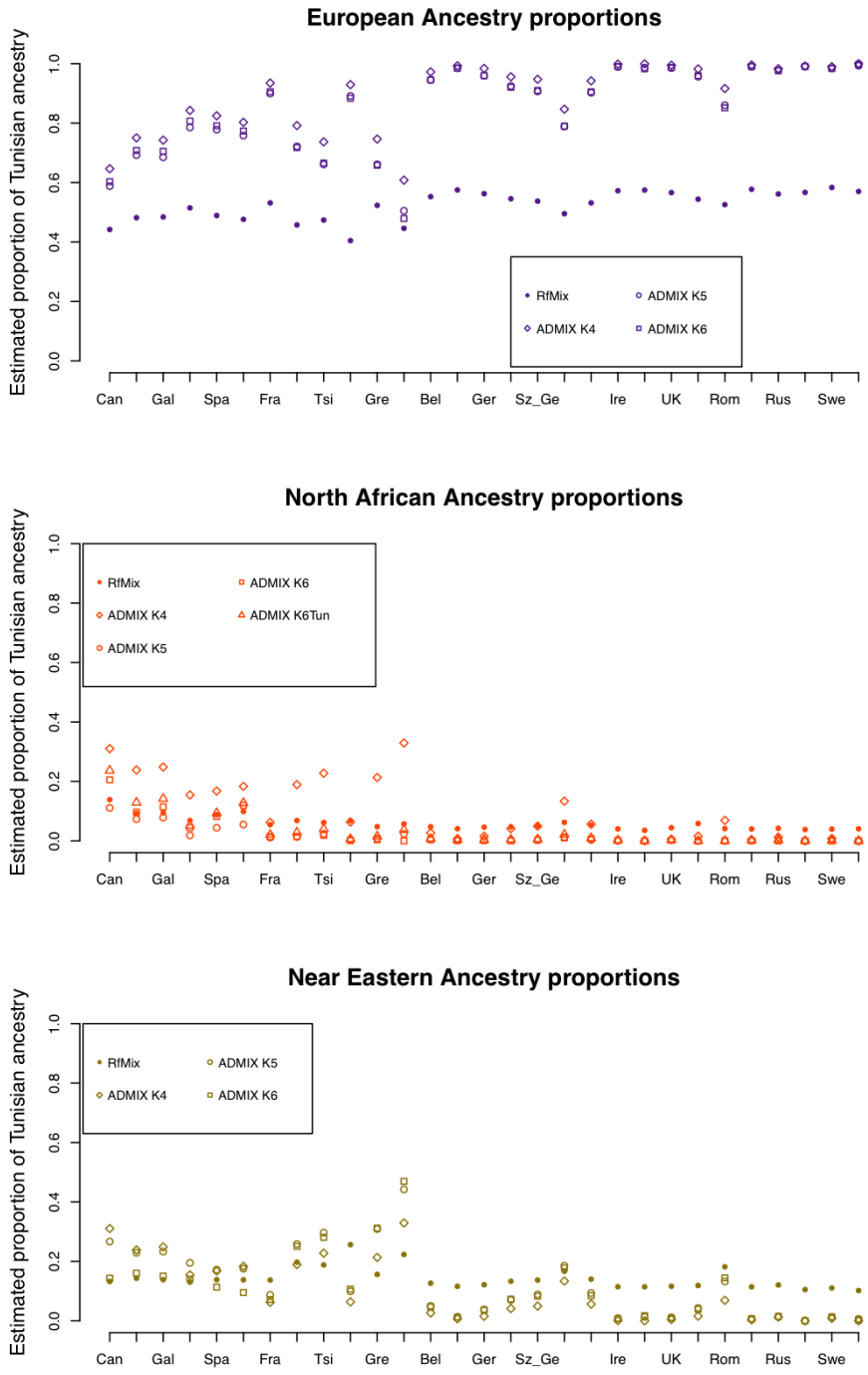

Figure S4: *Ancestry proportions estimated in European populations using RFMix and ADMIXTURE for k = 4, 5 and 6*. a) European ancestry b) North African ancestry c) Near Eastern ancestry. For RFMix only calls with more than 99% of confidence are shown, leaving an average of 30% of the genome uncalled but yet included in the denominator when estimating ancestry proportions. Results show that estimates of European ancestry follow a similar pattern using both methods, and the difference in proportions is most likely due to the uncalled ancestry, whereas for North African and Near Eastern ancestries, correspondence between both methods is highest for k = 5 and 6.

Figure S5

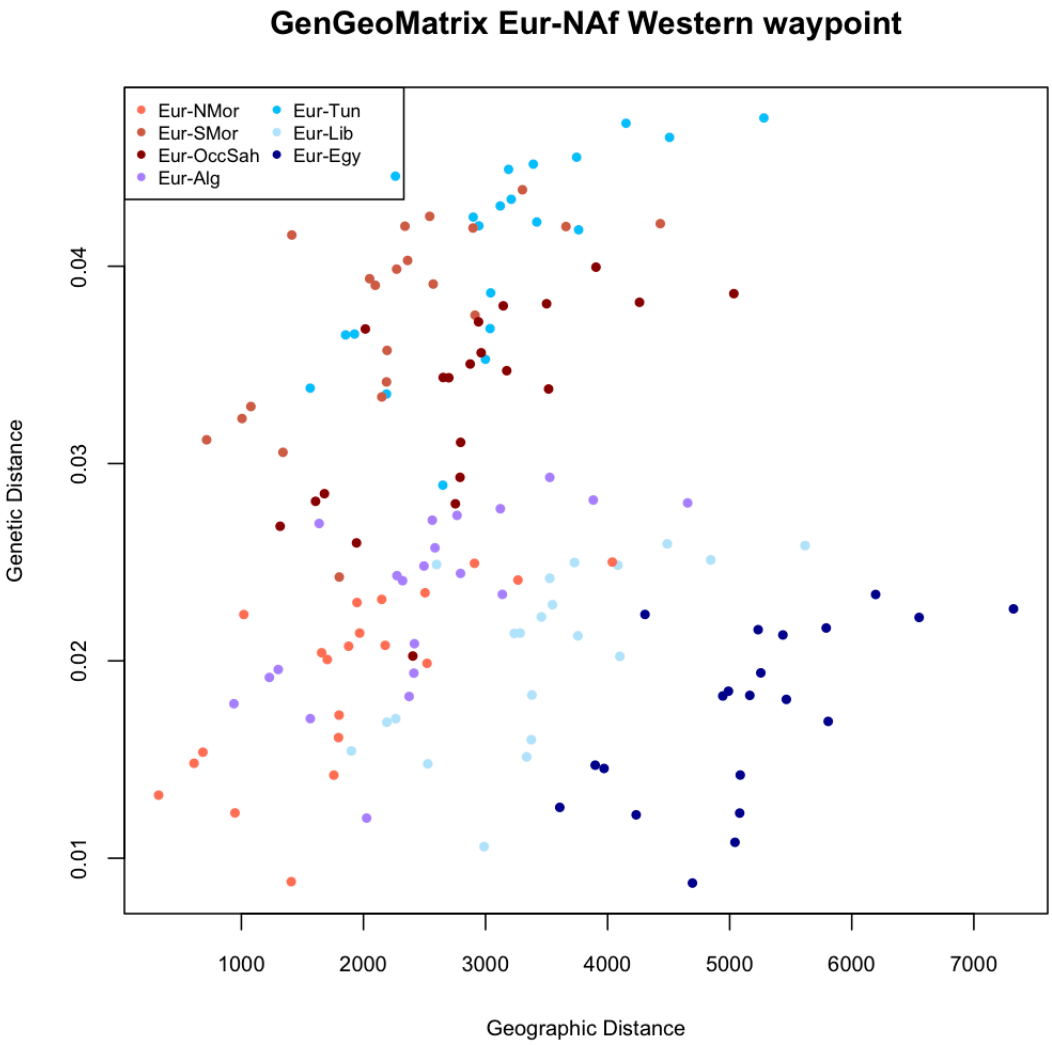

Figure S5*: Relationship between geographic (x-axis) and genetic (y-axis) distance between European and North African populations*. Results show a lack of correlation between the two parameters.

Figure S6

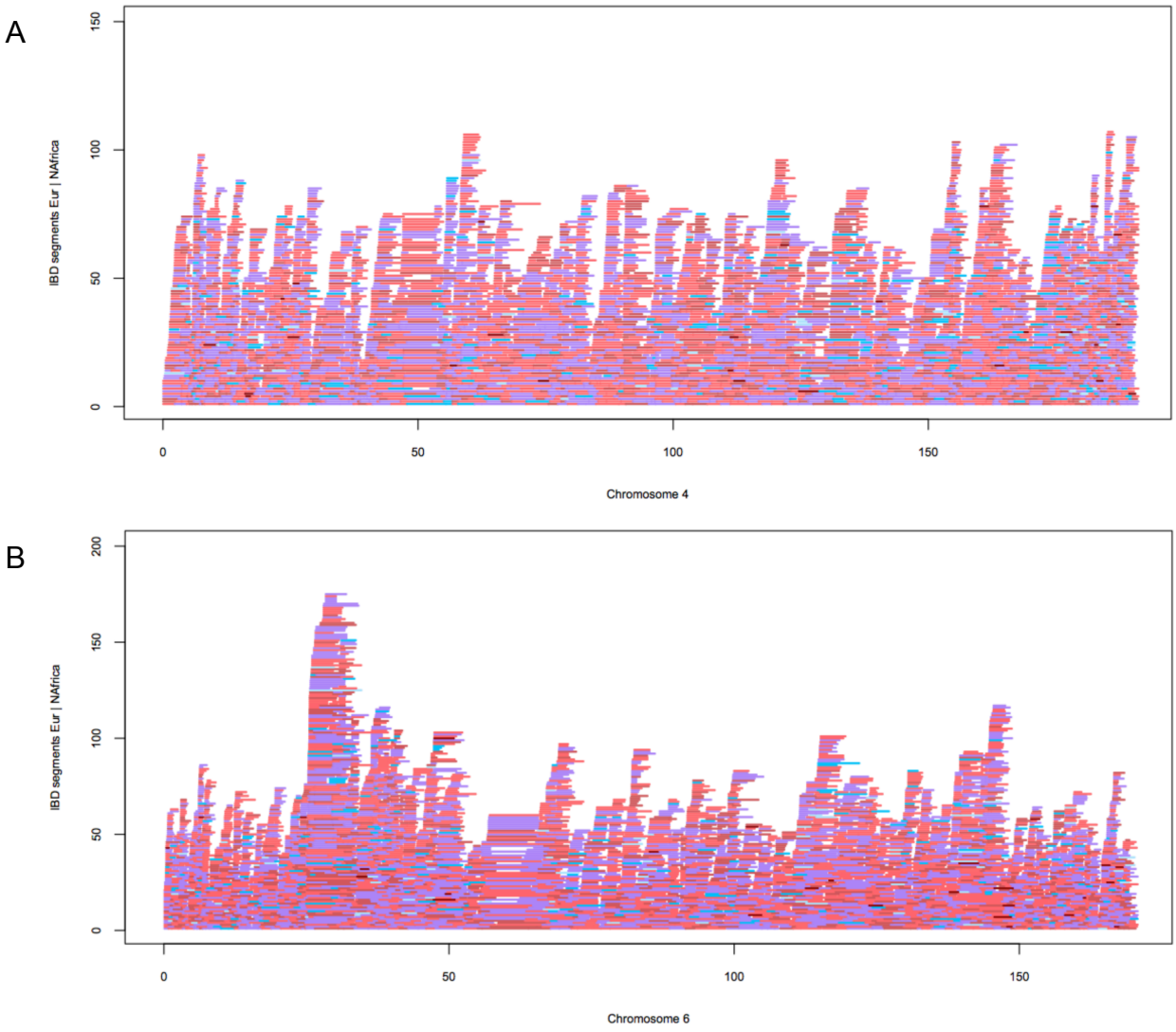

Figure S6*: Example of North Africa - European IBD distribution along a chromosome*. A. Results for chromosome 4 where no evidences of positive selection are observed. B. Results for chromosome 6 where peaks of IBD sharing are in agreement with a positive selection pressure in the HLA region. Different colors mean different locations within Europe.

Figure S7

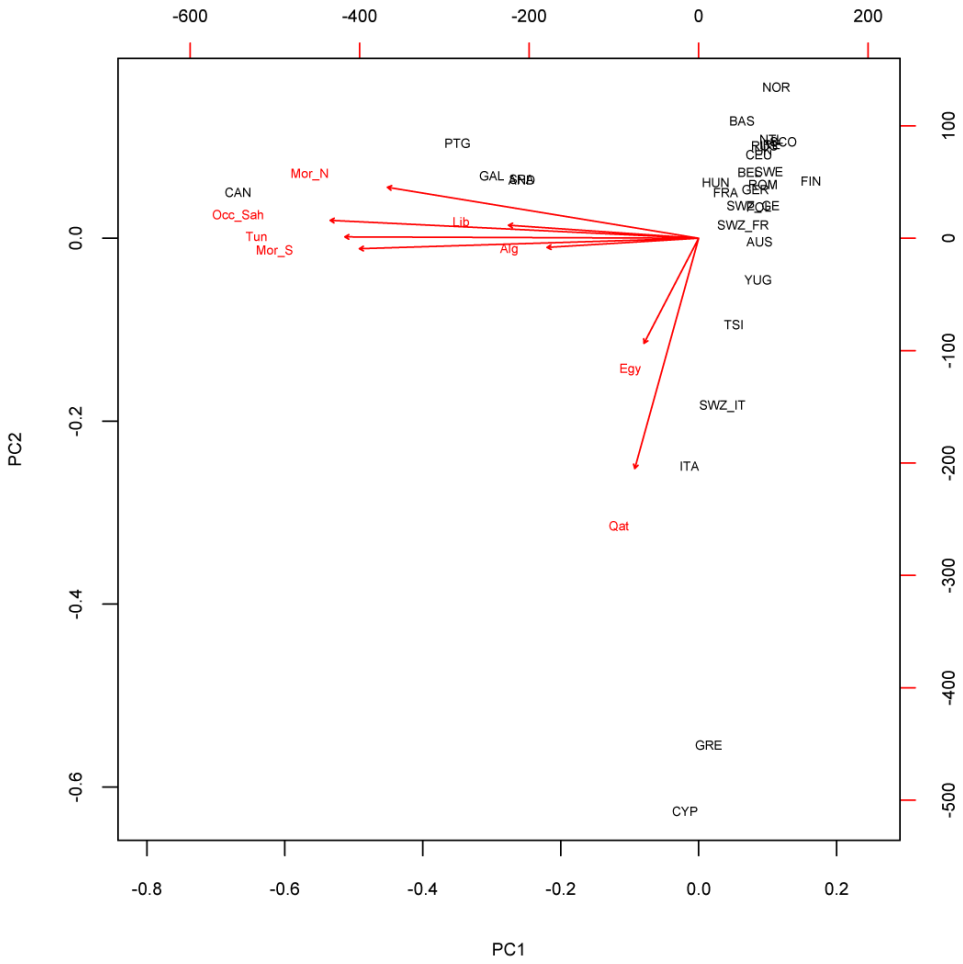

Figure S7*: Principal Component Analysis on European populations according to their* W_{EA} *values at the population level*. The contribution of each North African population and of Qatari to the overall analysis is represented in red arrows.

Figure S8

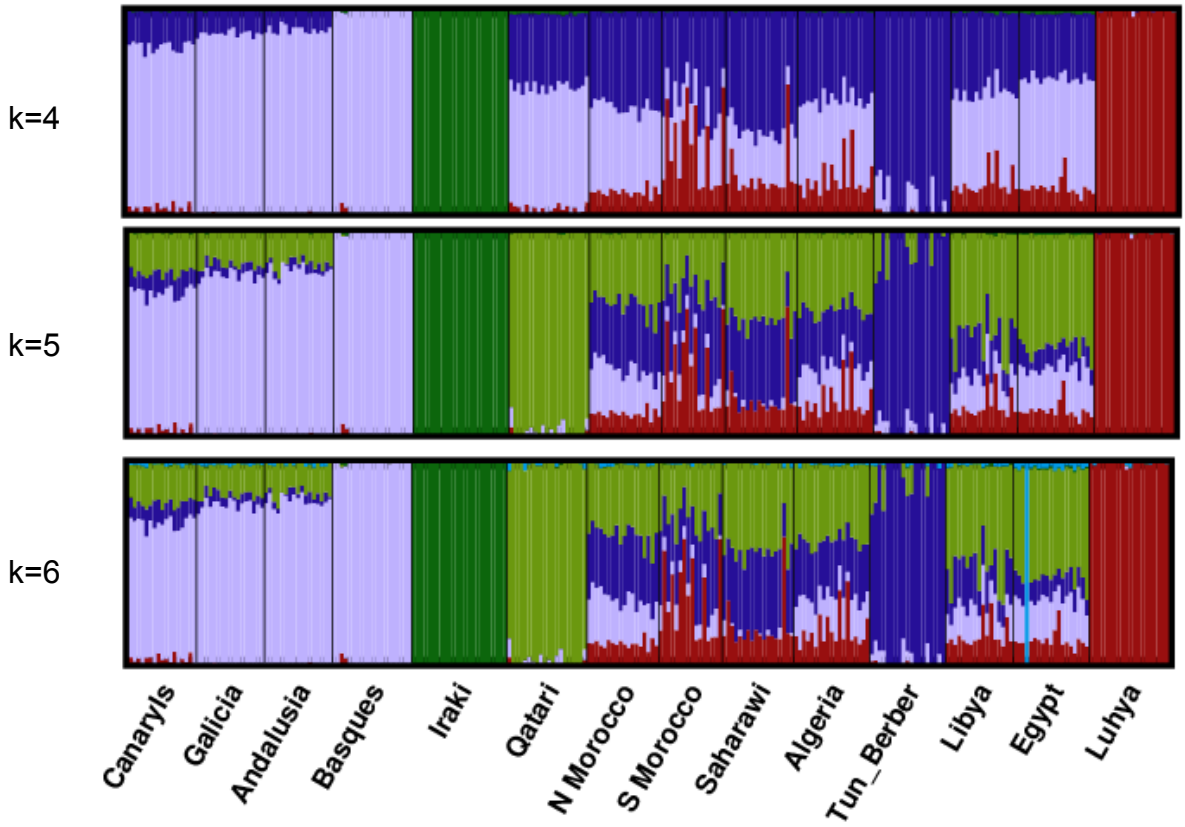

Figure S8: *Comparison of Near Eastern source populations. ADMIXTURE results for k= 4 through 6 from a dataset including populations from Europe, North Africa and the Near East*. Our goal was to ask whether the Iraqi (9) or the Qatari were better source populations of the Near Eastern ancestry found in Europe and North Africa. Results show that no allele sharing is detected between Iraqi and the other populations. However, the ancestral component assigned to the Qatari populations is present in both North African and European populations, indicating that the Qatari are a better genetic representative of the Near Eastern influence in other regions.

Figure S9

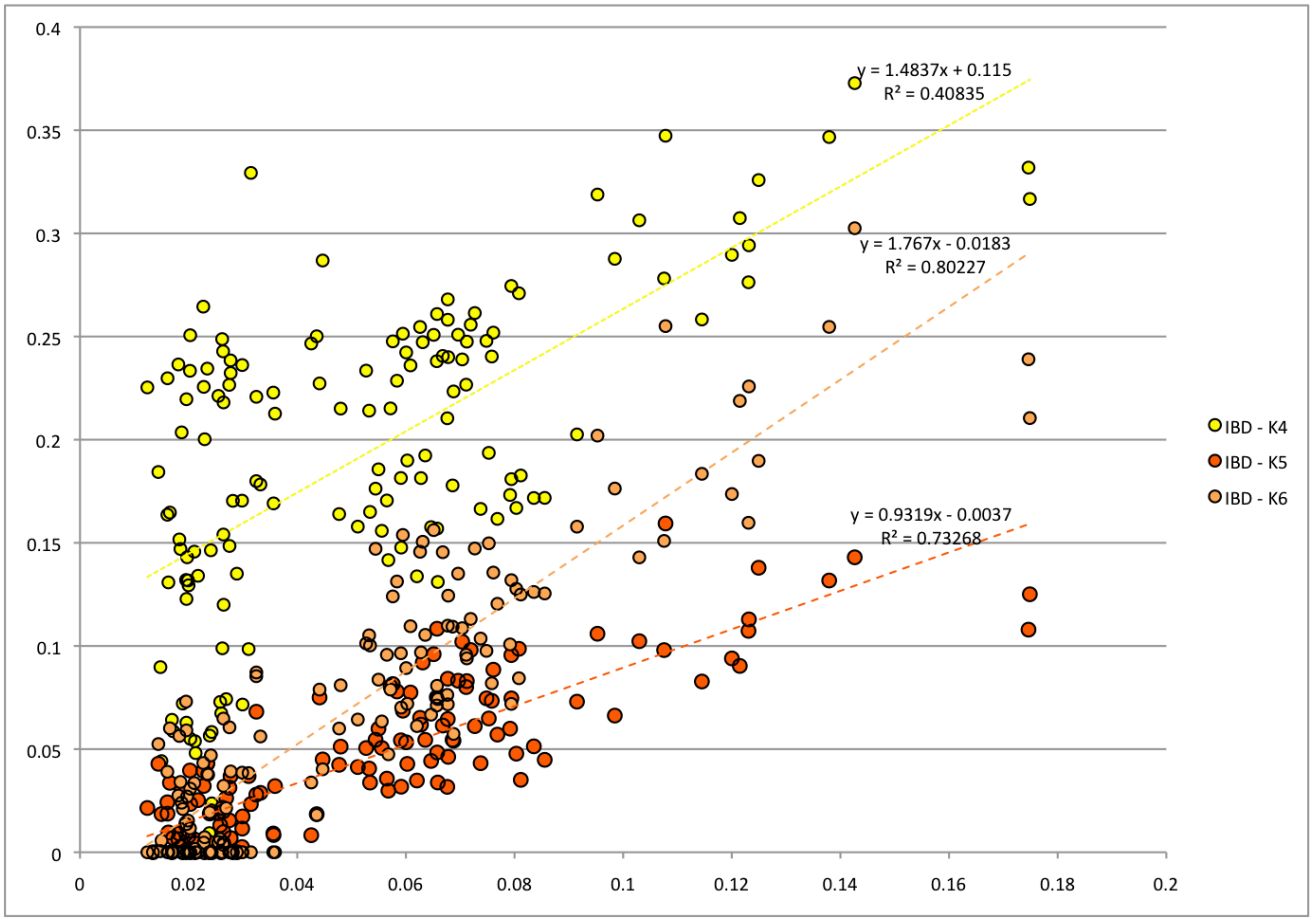

Figure S9*: Correlation between ADMIXTURE and fastIBD North African ancestry estimates*. Results show that a better correspondence exists between ancestry assignment with fastIBD and ADMIXTURE k=6 (R_2 = 0.86). However, the correlation between fastIBD and ADMIXTURE k=5 is the closest to 1 (slope 0.97).

Figure S10

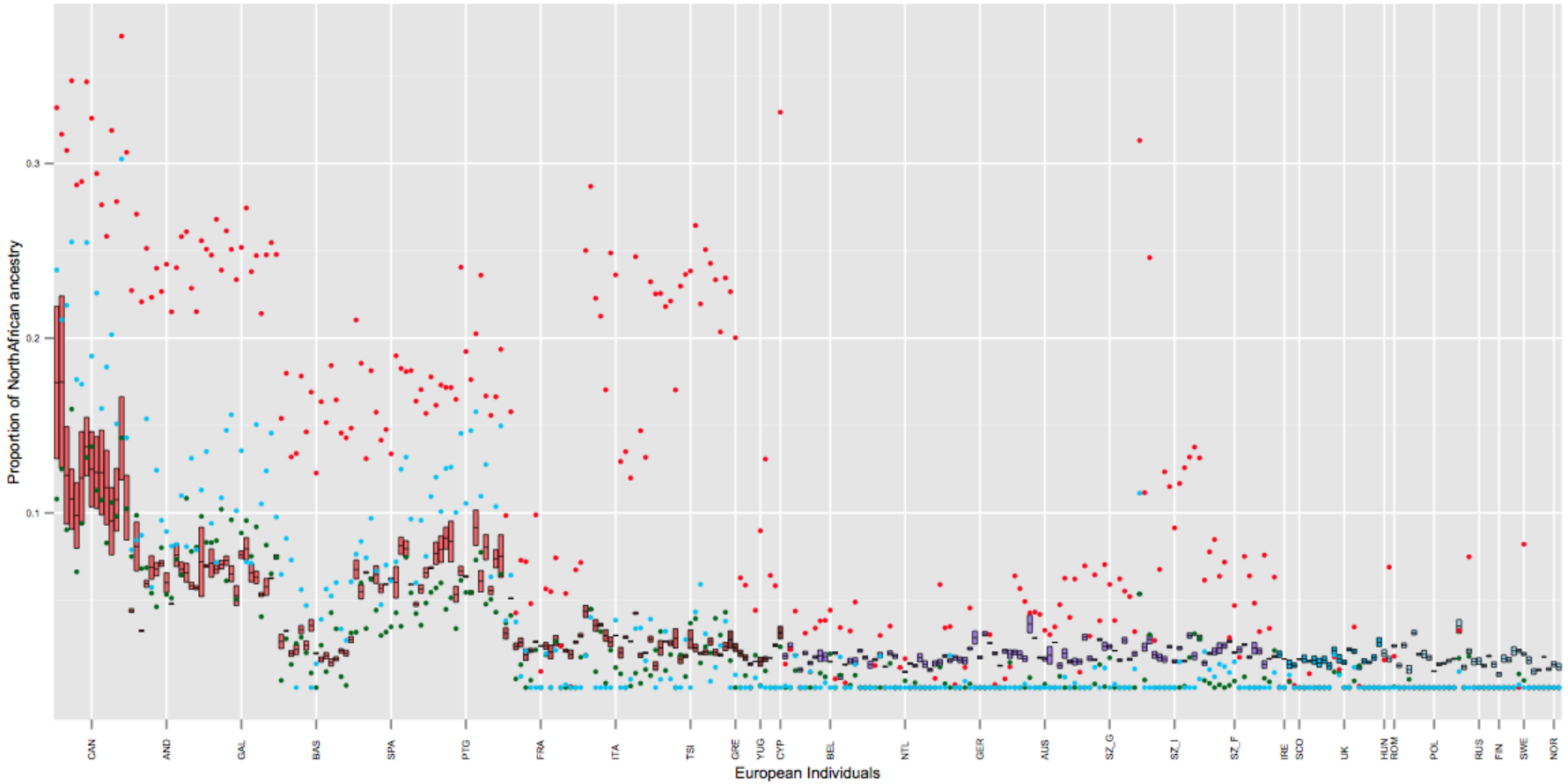

Figure S10: *Estimates of North African ancestry inferred using IBD and ADMIXTURE at individual level in Europe*. IBD estimates of ancestry are based in the number of shared segments between a given European individual and North Africa. fastIBD does not provide information of the phase of the shared segment between two given individuals. Thus, ancestry is represented as a *range*, being the lower bound the number of cM shared assuming that each repeated segment comes from the same phase (and thus is counted only once), and the upper bound the number of shared cM assuming that repeated segments come from different phases. ADMIXTURE ancestry point estimates are the ones for Near Eastern | North African ancestry and North African ancestry for k= 4 and 5, 6 respectively. Note that ancestry proportions detected by IBD sharing are similar to those reported by ADMIXTURE at k=5 and 6. Considering that the first method detects recent ancestry and the second more ancient one, we can assume that most of the IBD segments detected have a North African origin.

Figure S11

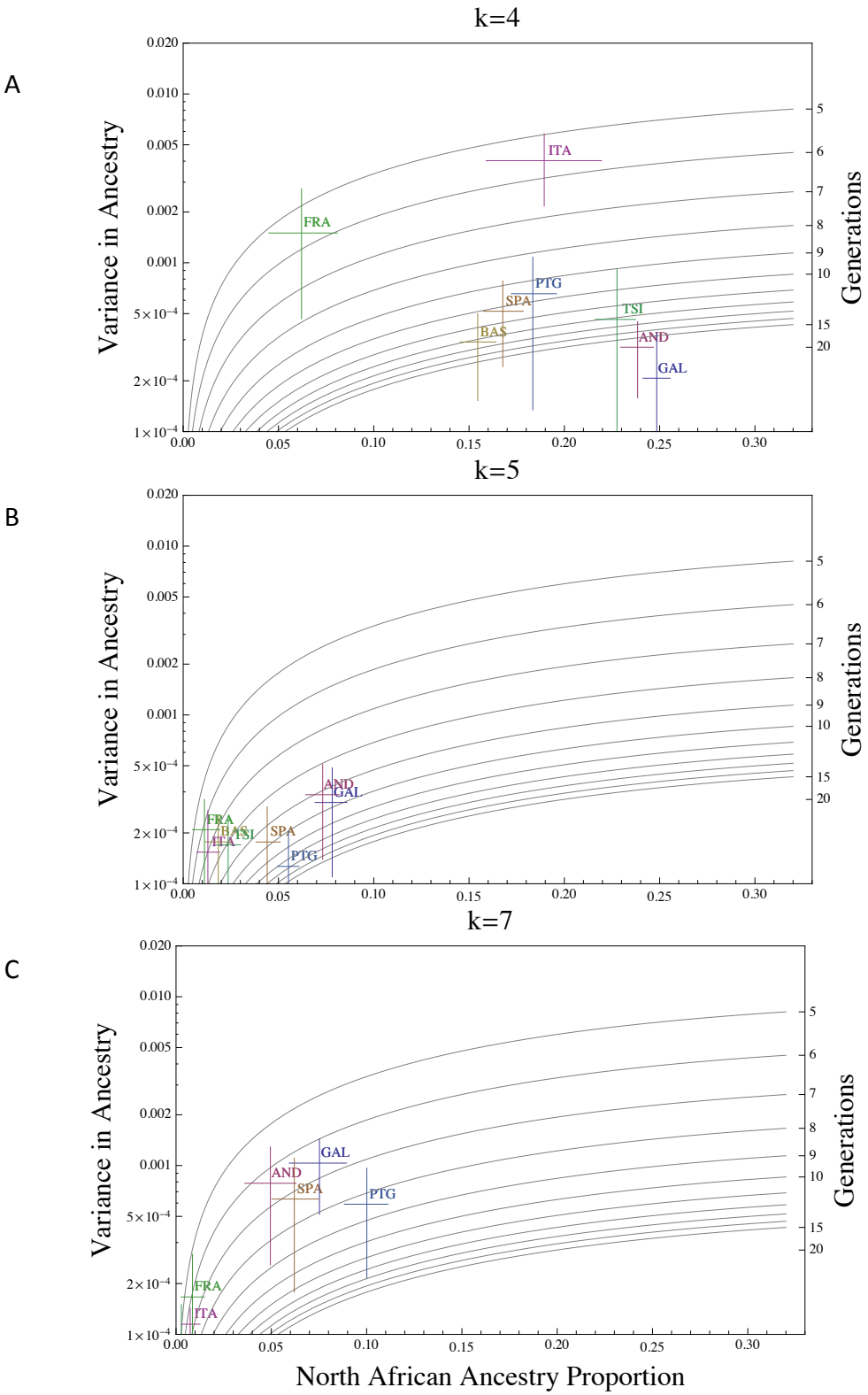

Figure S11: *Variance in ancestry proportions within populations depends on the overall ancestry proportions in the population and the time of gene flow*. A Estimating the effective time of migration based on variance in Near Eastern | North African ancestry proportions inferred under the k=4 model. Estimates of effective migration time based on North African ancestry proportions inferred under B k=5 model, C k=7 model.

Figure S12

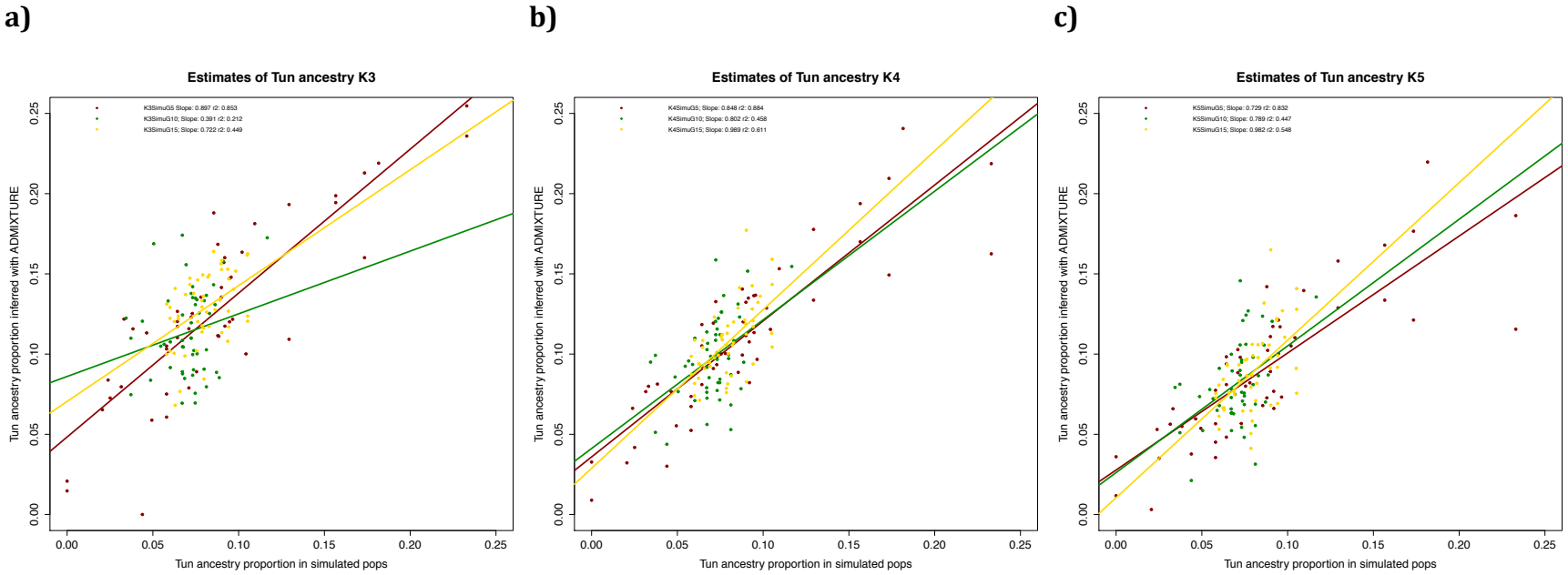

Figure S12: Correspondence between simulated Tunisian ancestry proportions 5, 10 and 15 generations after an admixture event and North African ancestry proportions inferred with ADMIXTURE. **a)** Results for $k = 3$, **b)** Results for $k = 4$, **c)** Results for $k = 5$. Correlation measures by r_2 is highest for $k = 4$ and for the 5 generation simulations.

Figure S13

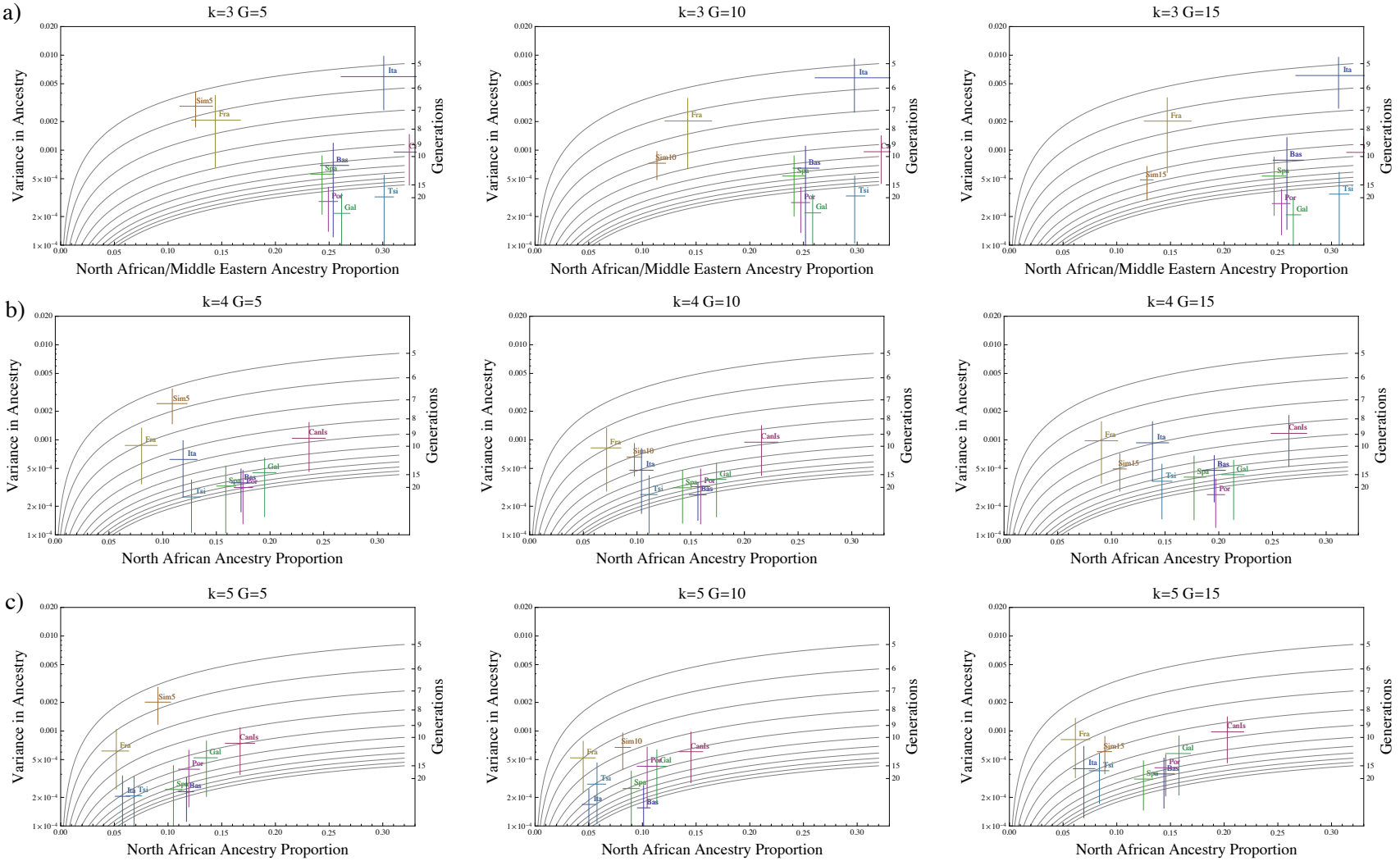

Figure S13: *Variance-based time since admixture estimates using a simulated, admixed European population for G = 5, 10, 15 generations since a pulse of North African migration into a European population*. The simulated, admixed population was then run with the original ADMIXTURE panel (minus the individuals used to seed the simulations) for **a)** $k = 3$, **b)** $k = 4$, **c)** $k = 5$. This allows us to both observe the sensitivity to the assumption of k, as well as the sensitivity to time since admixture. The proportion of admixed ancestry in the simulated population remained the same across runs, at 9% North African ancestry.

Figure S14

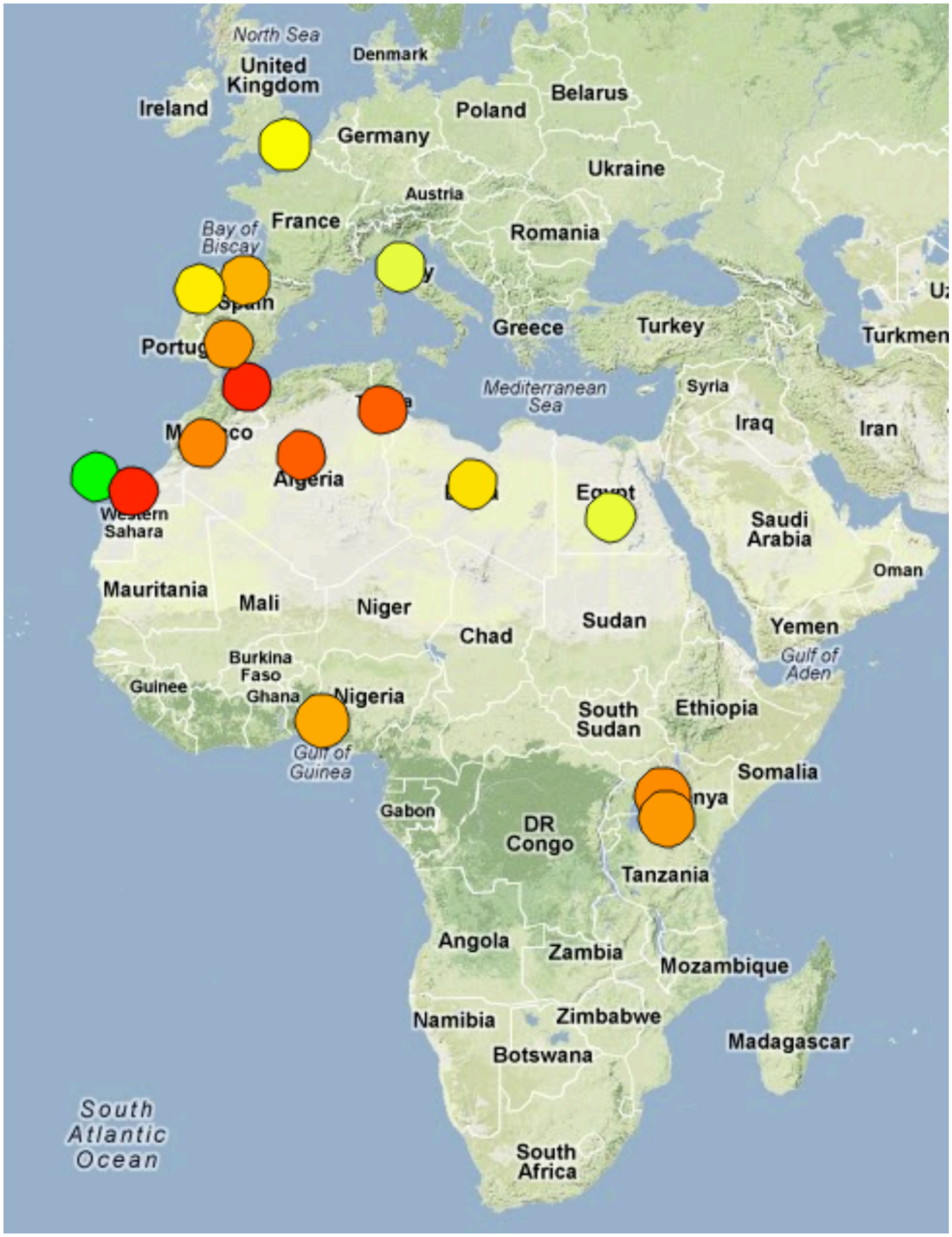

Figure S14 *Risk scores for multiple sclerosis in a set of Sub-Saharan, North African and European populations*. Map representing the cumulative risk allele frequency results for multiple sclerosis in different populations. Multiple sclerosis does not conform to the expected pattern of neutral drift for different populations, suggesting some effect of natural selection. Scores are represented as colored circles where green is the lowest cumulative risk allele frequencies (0.45) and red is the highest (0.55).

Figure S15

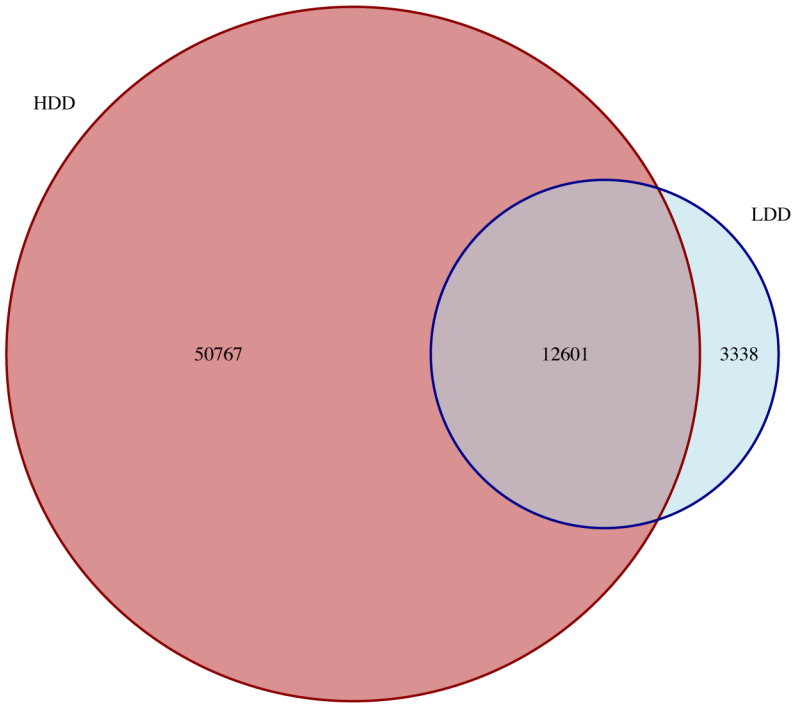

Figure S15 *Venn Diagram showing the proportion of LDD and HDD IBD segments detected within Europe that occur in both datasets*. Circles are proportional to the number of segments detected in each dataset. In red, segments detected in HDD are shown, and in blue segments detected in LDD. The overlapping region represents 80% of the LDD segments.

Figure S16

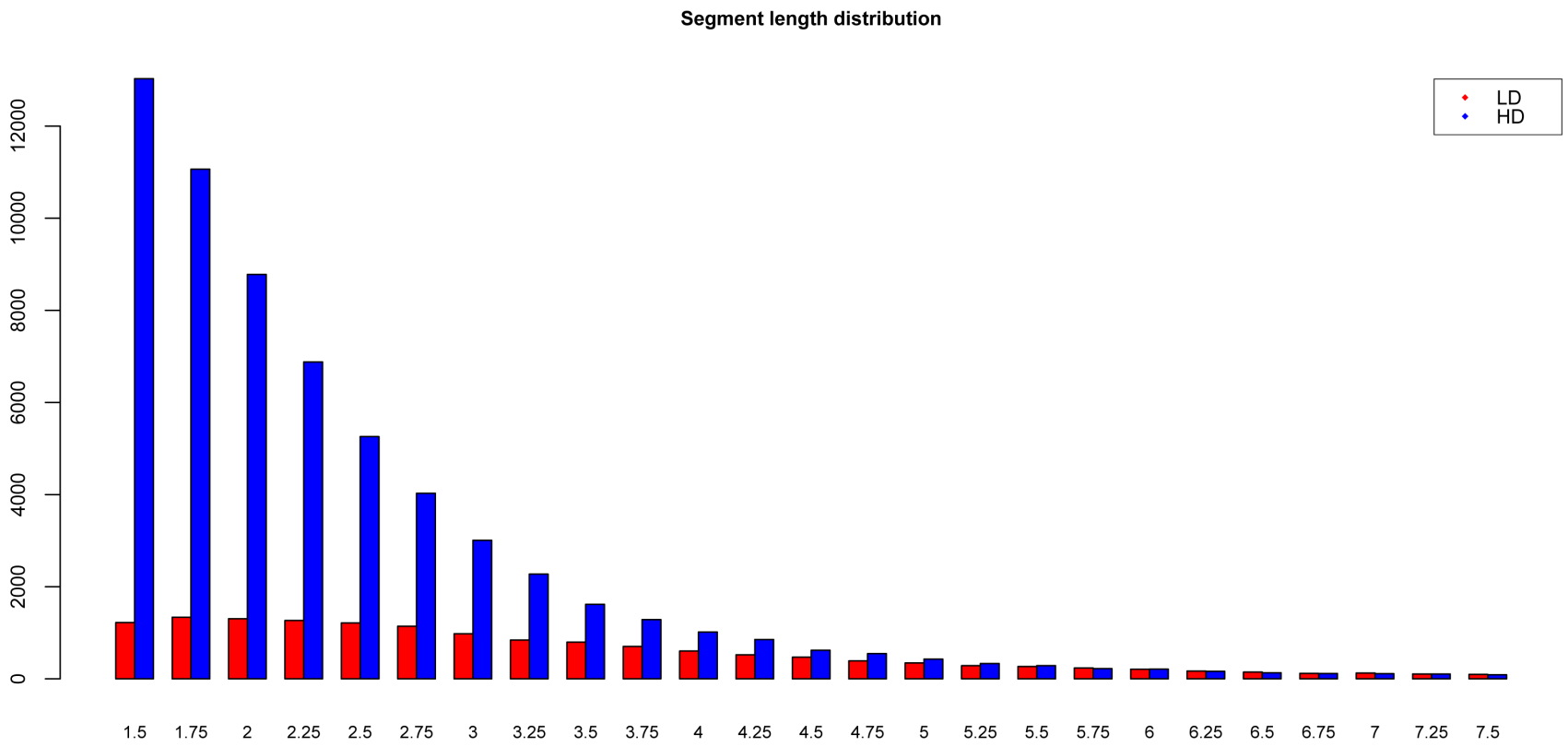

Figure S16: *IBD segment lengths inferred from high-density (HD) and low-density (LD) datasets. The length distribution of segments detected to be shared IBD within Europe*. Only segments between 1.5 and 7.5 cM are shown. Segments inferred to be IBD between North Africa and Europe in the HDD follow a clear exponential shape, and the HDD detects many more segments than LDD below 4 cM. Both patterns point to the fact that LDD is not able to detect all short segments. Oscillations in the distribution of segments shared between Europe and Sub-Saharan Africa might be explained by errors in the iterations of the phasing process that could alter the identification of shared segments.

Figure S17

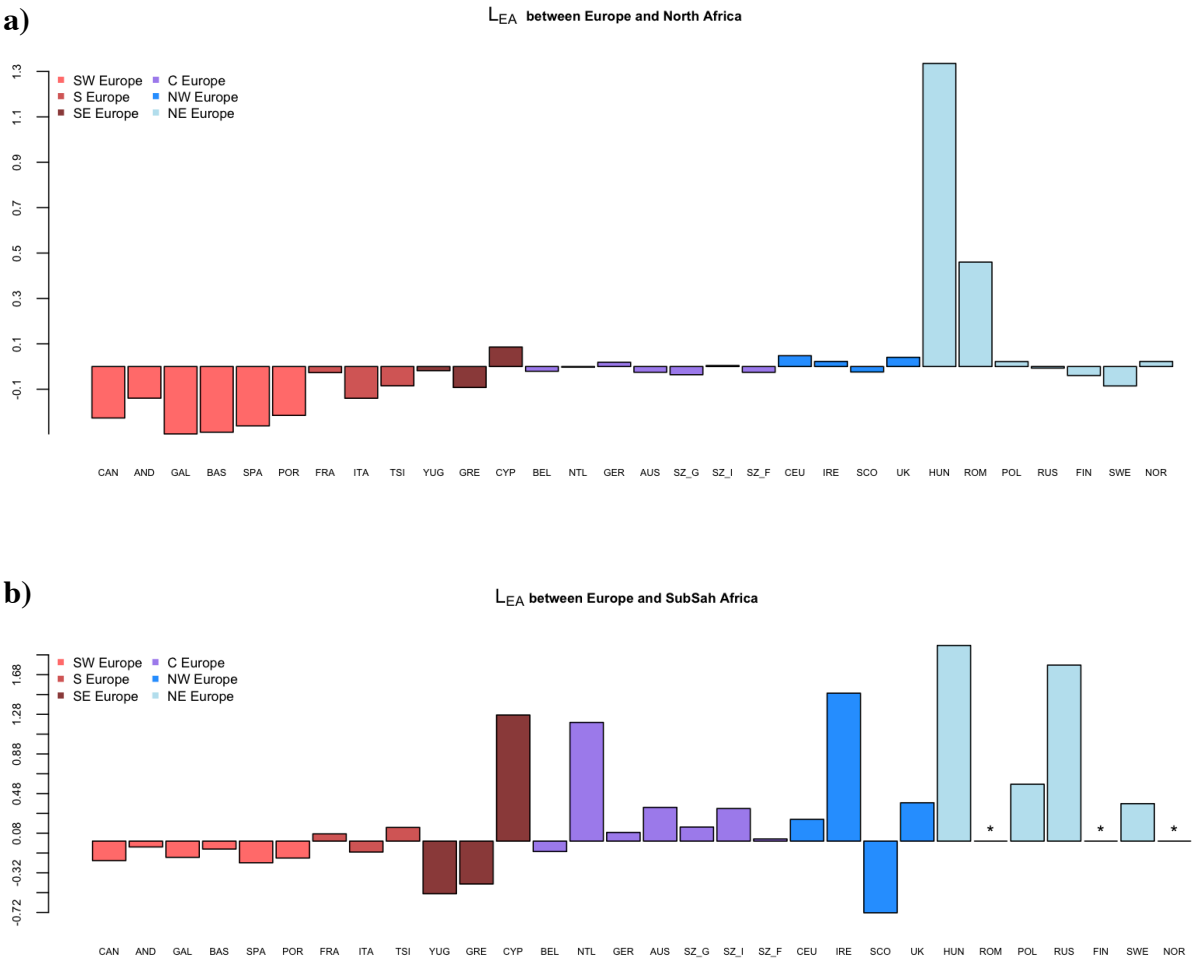

Figure S17: *Mean length estimates of IBD segments shared between Europe and Africa*. L_{EA} results between (a) Europe and North Africa and (b) Europe and Sub-Saharan Africa. The 0 value of the x-axis represents the mean length of a shared segment, and the bars show the deviation from this mean for each European population. An increasing geographic gradient of the length can be appreciated in the segments shared with North Africa, whereas the shared segments in Sub-Saharan Africa have a much bimodal shape, having South European populations a much shorter length than the mean, while the rest of the European populations have shared segments much longer than the European mean. A Mann-Whitney U-test using segments length between Europe | Sub-Saharan Africa and North Africa | Sub-Saharan Africa was performed to see if differences between the length distribution of the two groups existed. Results showed that segments length followed the same distribution (p-value = 0.2085). These different patterns between the African regions and the similar length distribution of Sub-Saharan shared segments including both North Africa and Europe are in agreement with a scenario characterized by extensive gene flow with North Africa to Europe and reduced gene flow with Sub- Saharan Africa through North African migrants. "*"indicates when no shared segments are found.

References:

1. Gusev A, *et al.* (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19(2):318-326.
2. Browning BL & Browning SR (2011) A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 88(2):173-182.
3. Albrechtsen A, Moltke I, & Nielsen R (2010) Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186(1):295-308.
4. Corona E, Dudley JT, & Butte AJ (2010) Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PLoS One* 5(8).
5. Henn BM, *et al.* (2012) Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *PLoS Genetics* 8(1):12.
6. Atzmon G, *et al.* (2010) Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am J Hum Genet* 86(6):850-859.
7. Lao O, *et al.* (2008) Correlation between genetic and geographic structure in Europe. *Curr Biol* 18(16):1241-1248.
8. Novembre J, *et al.* (2008) Genes mirror geography within Europe. *Nature* 456(7218):98-101.
9. Xing J, *et al.* (2010) Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 96(4):199-210.