

Supporting Information

African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*

Arun Durvasula*, Andrea Fulgione*, Rafal M. Gutaker, Selen Irez Alacakaptan, Pádraic J. Flood, Célia Neto, Takashi Tsuchimatsu, Hernán A. Burbano, F. Xavier Píco, Carlos Alonso-Blanco, Angela M. Hancock

SUPPLEMENTARY TEXT

Supplementary Methods

Samples. We sequenced the genomes of 79 *A. thaliana* individuals, including 70 fresh samples and 9 herbarium samples (Table S1). The majority of fresh samples were grown from seeds originally collected in Morocco, with 67 accessions collected across the Rif and Atlas mountains, including 64 from (1) as well as three additional samples (Aitba, Toufl, and Ita-0) that we obtained from the Nottingham Arabidopsis Stock Center (NASC). The 67 samples that were collected as part of a previous study had originally been genotyped at 249 SNPs as members of a panel of 151 Moroccan accessions (1). Here, we maximized variation by choosing for sequencing the subset that were not identical at the 249 SNPs previously genotyped. In addition, we sequenced two individuals from Mount Ketumbeine in Tanzania and one individual from Platres, Cyprus. The herbarium samples comprise one individual from Algeria, three from Tanzania (one from Mount Kilimanjaro and two from Mount Meru), and five from South Africa.

Sample collection, DNA extraction and DNA sequencing. The modern samples included in our study came from freshly grown leaf material. Seeds were stratified for four days and plants were grown under standard conditions in growth chambers. DNA was extracted from young leaves using ThermoScientific GeneJet Plant Genomic DNA kits and quantified using a Qubit Fluorometer, and quality was assessed using a fluorescence Nanodrop machine. Sequencing libraries were prepared using Illumina TruSeq DNA sample prep kits (Illumina, San Diego, CA) and sequenced on Illumina Hi-Seq instruments. Average coverage across samples after quality filtering is 27.2 (minimum=15.1; maximum=42.9) for the whole genome, and 31.8 (minimum=18.0; maximum=48.3) in the subset of the genome used for analyses (excluding missing data). Coverage details for all samples are provided in Table S1.

Herbarium specimens from South Africa were sent to us by curators at SANBI and sampled at the Department of Botany at the University of Vienna. Two plants from Mount Meru in Tanzania and one plant from Algeria were sampled at Naturalis Biodiversity Center (Leiden). DNA from herbarium specimens was extracted as described previously (2) in a clean room facility at the University of Tübingen. In short, for each sample, herbarium tissue was ground in a 2.0 ml tube with a metal pestle and incubated with PTB/DTT lysis buffer at 37°C overnight. DNA solution was transferred onto a QIAShredder column and from that point a DNEasy kit (Qiagen) protocol was followed. Two independent genomic libraries were constructed from 20 µl of DNA extract for each sample. These included (i) libraries without enzymatic

49 removal of cytosine to thymine (C-to-T) substitutions typical of ancient DNA (aDNA)
50 associated damage and (ii) libraries treated with uracil glycosylase (UDG), which
51 removes the excess C-to-T substitutions associated with damage. Libraries without
52 uracil glycosylase treatment were used to test the authenticity of the reads derived
53 from herbarium specimens, while uracil glycosylase-treated libraries are devoid of
54 aDNA-associated damage and were used for deep sequencing and subsequent analysis
55 (3). Shotgun libraries were constructed following a published protocol (4) with
56 modifications suggested in (5). Libraries were amplified for 10 cycles with unique
57 combinations of two indexing primers (6). The quality of each library was tested in
58 two consecutive RT-qPCR reactions (prior to and after the indexing amplification
59 step) and in a Bioanalyzer. Non-UDG treated libraries were sequenced on an Illumina
60 MiSeq instrument. UDG-treated libraries were modified by addition of USER™
61 enzyme at a blunting step and sequenced on the Illumina HiSeq 3000 platform.
62 Average coverage of herbarium samples after quality filtering is 7.5 (minimum=4.5;
63 maximum=9.1) for the whole genome, and 10.1 (minimum=6.8; maximum=12.1) for
64 the subset of the genome used for analyses, i.e., excluding missing data. Coverage for
65 individual samples is shown in Table S1.

66
67 **Authentication of herbarium DNA.** The main criterion for authentication of
68 ancient/historic DNA (aDNA) is the excess of C-to-T substitutions at the end of DNA
69 fragments, which is commonly referred as the aDNA damage pattern(7). We
70 constructed non-UDG-treated genomic libraries (as described above) in order to test if
71 this pattern is present in our historic samples. Additionally, with these libraries we
72 estimated the proportion of DNA molecules derived from *A. thaliana*, since in aDNA
73 samples some proportion of the reads are derived from microorganisms, which
74 commonly colonize historic samples. Sequenced reads were mapped to the *A. thaliana*
75 TAIR10 reference genome and the C-to-T substitution profile was calculated using
76 MapDamage v2.0 (8). The observed pattern in all our herbarium samples matched the
77 pattern of historic DNA (Fig. S13A), while the low levels of C-to-T substitution at the
78 first base (average 2%, minimum 1%, maximum 4%) were consistent with relatively
79 recent herbarium samples (9). Similarly, the distribution of DNA fragment lengths
80 (Fig. S13B), which we could estimate accurately in merged reads, is congruent with
81 time-dependent degradation (9). We estimated the proportion of *A. thaliana*
82 endogenous DNA through comparison of the number of reads successfully mapped to
83 TAIR10 reference from the total number of merged reads (Fig. S13C). The samples
84 had high levels of endogenous DNA (average 80%, minimum 62%, maximum 92%),
85 which allowed pure shotgun sequencing. Blank negative control samples, which were
86 amplified for 15 more cycles than test samples had much lower percentages of *A.*
87 *thaliana* endogenous DNA (Fig. S13). To control for possible cross-contamination
88 between herbarium samples that were processed together, we estimated levels of
89 heterozygosity in chloroplasts, which should be purely homozygous (excluding rare
90 instances of heteroplasmy). To that end, we called variants in chloroplast genomes
91 using the following steps in a GATK pipeline: HaplotypeCaller, GenotypeGVCF and
92 VariantFiltration (10). Our analysis revealed that the few potentially heterozygous
93 sites (proportion of homozygous calls < 0.8) in test samples were predominantly
94 indels with low coverage (Fig. S13D), which were likely the result of mismapping.
95 We concluded cross-contamination between our historic samples was negligible and
96 therefore unlikely to affect base calls given our filters.

97

98 **Alignment and SNP-calling.** For modern samples, we first trimmed adapters using
99 adapterremoval v2.1.2 (11) with parameters `<--trimns --trimqualities>` and picard-
100 tools v1.100(12) SamToFastq tool for conversion between .bam and .fastq file
101 formats. Sequencing reads for herbarium samples were trimmed using Skewer
102 v0.1.120 (13). Overlapping pairs of sequences in herbarium samples were merged
103 using Flash (14), which increases the quality of called bases in short molecules.
104 Herbarium samples were integrated into the analyses and SNPs were called in the
105 same way as modern samples, except that paired end reads were merged into single-
106 end reads and processed accordingly.

107 We used two different pipelines (described in detail below) to align sequences
108 to the Arabidopsis TAIR10 reference genome and to call variants. First, we used the
109 MPI-SHORE pipeline, already validated for processing *A. thaliana* re-sequenced
110 NGS data (15). In addition, we used a more conservative pipeline (FulgiPipe)
111 designed to reduce false positives due to indels, using a custom Java program. This
112 pipeline was necessary for population history reconstruction with MSMC because the
113 results of this method are strongly affected by linked errors. Results for other analyses
114 were highly similar between pipelines.

115
116 **Sequencing pipelines.** We used two different pipelines to align sequences to the
117 Arabidopsis TAIR10 reference genome and to call variants. First, we used the MPI-
118 SHORE pipeline, already validated for processing *A. thaliana* re-sequenced NGS data
119 (15). In addition, we developed a second, more conservative, pipeline (FulgiPipe) to
120 be used for MSMC analyses, where linked errors affect the results.

121
122 More specifically, the MPI-SHORE pipeline used the following software and
123 parameter settings: First, to pre-process the TAIR10 reference genome we used the
124 MPI-SHORE subprogram 'preprocess' with parameters `<-C --indexes`
125 `BWA,SuffixArray>` as well as the bwa v0.7.5a (16) command 'index' with parameters
126 `<-a bwts>`. We further used this software for alignment of the reads to the reference
127 genome, using the commands 'aln' with parameters `<-n 0.1>` and sampe with
128 parameter `<-a 500>` (samse for herbarium samples). Then we imported trimmed fastq
129 files in shore format using the MPI-SHORE subprogram 'import' with parameters `<--`
130 `application genomic --importer Fastq --shore-filter --max-Ns 10% --lowcomplexity>`.
131 Various file conversions relied on MPI-SHORE subprograms 'convert
132 Alignment2Maplist' and 'convert Variant2VCF'. Finally, variants were called with
133 the MPI-SHORE subprogram 'consensus' using the empirical scoring matrix
134 approach and parameters `<-b 0.9 -g 4 -h 6 -i 0.5 -N>`.

135
136 The more conservative pipeline was designed to reduce false positives due to indels,
137 using a custom Java program. Specifically, we excluded low complexity regions of
138 the genome, where alignment of reads is challenging and prone to error. In particular,
139 we classified as missing data all regions where the same base is repeated five or more
140 times in the reference genome, as well as the adjacent ten bases. In addition, we
141 excluded the first and last positions of each read, we filtered to remove bases with
142 quality < 30 and coverage $< 5x$, and we eliminated positions with coverage greater
143 than twice average coverage to remove potential duplications in the sample not
144 represented in the reference genome. For variant calling, a calling ratio threshold of
145 0.0 to 0.2 was used to call a reference allele, and of 0.8 to 1.0 to call a mismatch to the
146 reference. Further, to avoid strand-specific errors, mismatches were called only if they
147 were supported by at least one read aligned on both the forward and reverse strand. In

148 case the calling ratio was consistent with a mismatch, but only reads on one or the
149 other strand supported the call, the position was recorded as missing data. This
150 pipeline was used for population history reconstruction with MSMC because the
151 results are strongly affected by linked errors. Results for other analyses were highly
152 congruent between pipelines.

153
154 **Quality Control and Error Rate Estimates.** To estimate the error rate in
155 sequencing, mapping and variant calling, we independently sowed, grew and
156 sequenced two biological replicates of the same Moroccan accession (Ket10), plus
157 four replicates of the European accession Ma-0. In total, we compared the sequences
158 of seven pairs of putatively identical accessions. Each sample was independently
159 mapped and variants were called with our two pipelines, and the genomes of each pair
160 were compared in terms of number of differences, and in terms of base pairs called as
161 non-missing data (Table S6). Assuming no residual heterozygosity in the parents and
162 a mutation rate of $(7.1 \pm 0.7) \times 10^{-9}$ mutations per base pair per generation (17) the
163 expected rate of real differences between pairs is $(1.42 \pm 0.14) \times 10^{-8}$ per base pair.

164
165 The MPI-SHORE pipeline identified on average 14251.6 differences between pairs of
166 identical samples, while FulgiPipe called on average only 8 differences (Table S6).
167 Overall, these results show that the more conservative pipeline has a much lower rate
168 of false variant calls (error rate= 9.9×10^{-8} per base pair), at the cost of excluding a
169 higher proportion of the genome from the analyses. In particular, FulgiPipe is
170 designed to remove linked errors, which could cause problems in MSMC analyses.
171 We therefore used the MPI-SHORE pipeline (error rate= 1.4×10^{-4} per base pair) for
172 analyses less sensitive to false variant calls, and FulgiPipe for MSMC. The two
173 pipelines resulted in similar patterns in analyses for which MPI-SHORE results are
174 presented.

175
176 We note, moreover, that samples did not cluster by sequencing lane, nor by
177 sequencing technology used. Herbarium samples, although necessarily sequenced and
178 processed with a different procedure, did not form a separate clade (Figs. 1 *B-C*), but
179 rather clustered by region of origin (Tanzania and South Africa, and Algeria), together
180 with modern samples from the same region.

181
182 **Effect of differences in coverage on diversity:** To check whether genomic variation
183 patterns are confounded by differences in coverage depth, we computed average
184 coverage and diversity (average pairwise differences, θ_{π}) for each of the nine Eurasian
185 and four Moroccan clusters. Moreover, we randomly subsampled reads in the raw
186 data of all African samples, to the minimum average coverage across Eurasian
187 clusters (15x, Fig. S7A), and repeated some of the analyses with this subsampled set.
188 For this purpose we used the samtools 1.3.1 (12) <view -s> function, and we
189 reprocessed all samples through the pipeline MPI-SHORE with the same parameters
190 described above.

191 There was no significant correlation between average coverage and diversity,
192 considering Eurasian clusters alone (Spearman rank correlation, $r_s = -0.25$, $p = 0.5165$),
193 Eurasian clusters together with Moroccan clusters ($r_s = -0.27$, $p = 0.3737$), and Eurasian
194 clusters together with subsampled Moroccan clusters ($r_s = -0.16$, $p = 0.5916$).

195 To check the effect of coverage depth at the level of single pairwise comparisons, we
196 computed pairwise differences per base pair (diff./bp) for every pair of African
197 samples (Fig. 2B) and for the same pairs after subsampling reads (Fig. S7C).

198 Artificially reducing coverage significantly lowered pairwise differences per base pair
199 (paired $t(1829) = 127.0$, $p < 0.01$). Nonetheless, this effect is of negligible magnitude
200 (mean difference = 0.00026 diff./bp, and see Fig. S7B) compared to the actual
201 differences across clusters (Fig. S7A, and compare Fig. 2B with Fig. S7C), so that
202 divergence among African groups after reducing coverage (range: 0.361-0.737 %
203 diff./bp, Fig. S7C) is of the same magnitude as before subsampling (range: 0.379-
204 0.767 diff./bp %, Fig. 2B). Even after artificially reducing coverage, Moroccan
205 clusters are as diverged from each other as Eurasian relicts from non-relicts (range:
206 0.356-0.748 % diff./bp).

207

208 **Population structure.** For population structure analyses, we subsampled the
209 complete data set to match sample sizes across clusters as some Eurasian geographic
210 regions are heavily oversampled, which could cause biases in some analyses (18-20).
211 To this end, we randomly selected seven samples from each of the Eurasian clusters
212 (defined as described above) as well as for each of the four Moroccan sub-regions. So
213 that no single region would drive the results, we pre-processed the data to remove
214 SNPs in strong LD using PLINK (21) --indep-pairwise 50 10 0.1 and we removed
215 SNPs with missing data by setting --geno 0, which retained 4,198,821 SNPs in the
216 PCA and NJ analyses and 4,818,354 SNPs in the ADMIXTURE analysis, which used
217 a different set of individuals (see ADMIXTURE analysis section).

218 We created a whole genome neighbor joining tree in R using the packages
219 adegenet v2.0.1 (22, 23), and ape v3.5 (24) both on the reduced set of samples (Fig.
220 1B) and on the complete set of all samples (Fig. S1). We performed principal
221 components analysis using the --pca option in PLINK (21) on the set of SNPs
222 described above.

223 We used the ADMIXTURE software (25) to cluster samples. We removed
224 populations represented by single samples because the software assumes population-
225 level data. The number of clusters (K) was determined based on the outcome of cross-
226 validation analyses (Table S2). For this, we ran ADMIXTURE twenty times and
227 calculated the mean cross-validation error for each K across runs (Table S2). Then,
228 we selected the K with the lowest mean cross-validation error (K=4). We also plotted
229 ADMIXTURE results with values of K ranging from 2-15 (Fig. S4). The analyses
230 based on ADMIXTURE were repeated with samples from Tanzania and South Africa
231 (Fig. S5). Although individually these regions would not reach the sample size used
232 for the rest of the clusters, we considered that the similarity across sub-Saharan
233 samples could provide a means to merge them.

234

235 In addition, to gain a better understanding of structure and history within Morocco, we
236 conducted PCA, ADMIXTURE and haplotype sharing analyses within this region
237 (Fig. S2). PCA and ADMIXTURE analyses were conducted as described in the
238 Methods section. For haplotype sharing analyses, we determined the extent to which
239 individuals shared DNA segments identical-by-descent (IBD) using the RefinedIBD
240 algorithm in Beagle v4 (26). We used the default parameters implemented in
241 BEAGLE and ran it on the samples from the 4 Moroccan regions.

242

243 **Pairwise differences.** We calculated per base pairwise differences between all pairs
244 of samples from Eurasia and Africa. As expected, the distribution of pairwise
245 differences within Eurasia (Fig. 2A) matches previous findings (see Figure 3A in
246 (27)).

247 To define a cutoff for calling relicts based on published results (27), we
248 calculated average pairwise differences between each previously categorized Eurasian
249 relict and all non-relicts. These ranged from 0.0050 to 0.0068 differences per base
250 pair (Fig. S6). We then classified any sample from our newly sequenced set as a relict
251 if the average number of pairwise differences was greater than the minimum for the
252 Eurasian relict/non-relict comparison (0.0050). This new set of relicts included all
253 Africans as we found that each African accession was more diverged from Eurasian
254 non-relicts than the relict with the lowest divergence. The range for African
255 accessions compared to Eurasian non-relicts is shown in Fig. S6.

256
257 **Geographic distances among clusters.** We compared the geographic distances
258 among clusters of Eurasian and Moroccan samples based on geographic coordinates.
259 For this, we used the function ‘distGeo’ in the R package geosphere v1.5-5 (28) to
260 compute for every pair of samples belonging to different clusters the geodesic
261 distance (i.e., shortest distance between their position on an ellipsoid with major
262 radius of 6378137 m at the equator and ellipsoid flattening $f = 1/298.257223563$,
263 consistent with the standard World Geodetic System WGS84). For every cluster, we
264 computed the median of the geodesic distances between samples of the focus cluster
265 and all samples belonging to other clusters, analyzing separately Eurasian and
266 Moroccan groups.

267
268 **Private variation.** We also used private variation as a measure of diversity. We
269 compared the nine Eurasian clusters to the Moroccans as a group (Fig. S8) as well as
270 to individual Moroccan sub-groups (Fig. 2C) using the following three measures:

- 271
272 1. **Private SNPs.** We counted the number of positions in the genome where one
273 of the alleles is present exclusively in the samples belonging to a single group
274 (denoted by ‘all SNPs’ in Figs. 2C and S8).
275 2. **Private SNPs, no singletons.** Since singletons represent the most external
276 branches in gene genealogies, they are strongly influenced by recent
277 population growth. Therefore, to better capture historical variation, we also
278 calculated the number of private SNPs as described above, but excluded
279 singletons. (denoted by ‘SNPs no singletons’ in Figs. 2C and S8).
280 3. **SNPs in private haplotypes.** Private SNPs that arose recently in the
281 population are highly unlikely to be linked to other private SNPs. Conversely,
282 the observation of several, contiguous private SNPs indicates ancestral
283 haplotypes not represented in other groups. Therefore, for the third measure of
284 private diversity, we restricted our analysis to stretches of contiguous private
285 SNPs (of length ≥ 2 SNPs), again after removing singletons (denoted by
286 ‘Haplotype Diversity’ in Figs. 2C and S8).

287
288 When calculating these statistics, we subsampled equally across clusters and
289 resampled 500 times within each cluster to avoid biases due to sample size
290 differences. The number of samples taken in each replicate differed between the
291 analyses in which Morocco was considered as a single cluster (where 20 individuals
292 were chosen per sample, Fig. S8) and when Morocco was separated into sub-clusters
293 (where 5 individuals were chosen per sample, Fig. 2C). This difference was due to the
294 sample sizes of the population with the lowest number of samples in each case (min.
295 sample size in the first configuration: Eurasian relicts with 25 samples; in the second
296 configuration: Rif population with 7 samples). Point estimates reported in Figs. 2C

297 and S8 are based on the mean across resampling, and 95% confidence intervals are
298 derived from the distribution across the 500 sampling iterations ($\pm 1.96 * SEM$).

299

300 **S-Locus analysis.** We downloaded S-locus reference sequences based on the
301 following sources: The S-locus haplogroup A reference derives from the Col-0
302 reference, the haplogroup B reference derives from Cvi-0 (29) and the haplogroup C
303 reference from LZ-0 (30). We trimmed adapters and aligned reads to a reference
304 created with these sequences using the same procedure described above for the MPI-
305 SHORE pipeline. We excluded reads with mapping quality less than 25 and assigned
306 the S-locus haplogroup that had the highest proportion of sites with non-zero coverage
307 to all individuals in our sample and in the worldwide set. We excluded samples from
308 the analysis that had less than 40% coverage for any S-locus reference, representing
309 overall low quality and/or low coverage samples. We validated our assignment of
310 accessions to S-haplogroups using the reference samples for S-haplogroups when they
311 were included in our sequence set (i.e., for accessions Col-0 and Cvi-0) as well as
312 other individuals for which haplogroups were called previously (C24, Kas-2, Br-0,
313 Pro-0, Ra-0, Mr-0 and Bur-0) (31).

314

315 **Demographic Inference with MSMC.** We used a sequentially Markovian coalescent
316 approach (MSMC v2 (32)) to infer patterns of effective population size over time. For
317 these analyses, we used the SNP calls from the conservative method described above
318 (FulgiPipe), because MSMC is very sensitive to clustered errors, such as those
319 resulting from improperly called SNPs around indels. Since *A. thaliana* accessions are
320 inbred, we created pseudo-diploids by combining chromosomes from pairs of
321 individuals from the same populations and ran MSMC on all pairwise comparisons
322 (Fig. 3A) Medians were fit with a cubic spline in Python and plotted \pm one standard
323 deviation shaded. We also plotted the medians of Eurasian clusters separately (Fig.
324 S10). In addition, we ran MSMC using 8-haplotypes (Fig. 3B), which provides greater
325 resolution on recent events.

326 We assumed a mutation rate of 7.1×10^{-9} based on results of mutation
327 accumulation experiments (17) and a generation time of 1 year (Fig. 3).

328 Cross coalescence rates were calculated across pairs of samples from the
329 populations of interest using the -P 0,0,1,1 option for 4 haplotypes in MSMC.

330 Relative cross coalescence rates were calculated as

331

332
$$2 * \frac{\lambda_{01}}{\lambda_{00} + \lambda_{11}},$$

333

334 where λ_{00} and λ_{11} represent coalescent rates within the two focal populations and λ_{01}
335 the coalescence rate across populations.

336 MSMC analyses have been shown with simulations to be unbiased in specific time
337 frames depending on the number of haplotypes used. Specifically, using parameters
338 appropriate for humans (generation time of 30 years), the method is reliable between
339 50 kya and 2 mya when using 2 haplotypes, and as recent as 2 kya using 8 haplotypes
340 (32). To translate these thresholds for *A. thaliana* we considered a generation time of
341 1 year, and a rate of effective outcrossing in natural environments of 4.1% (33). Since
342 MSMC relies on haplotype information for the inference, we considered that the
343 crucial parameter is the “effective generation time”, or the expected time to the next
344 outcrossing event, 1 year/generation x 100/4.1 generations/outcrossing event \sim 24.4
345 years/effective generation. Therefore, we considered that the thresholds of reliability

346 for MSMC in *A. thaliana* are for 2 haplotypes, 50 kya x 24.4/30 ~ 40 kya and 2 mya x
347 24.4./30 ~ 1.6 mya, and the estimates should be unbiased as recent as 2 kya x 24.4/30
348 ~ 1.6 kya using 8 haplotypes.

349
350

351 **Demographic inference with $\delta a \delta i$.** In order to confirm our demographic estimates,
352 we used $\delta a \delta i$ (34) to fit a simple population split model to the observed joint allele
353 frequency spectrum of different pairs of populations. For this analysis, we used only
354 intergenic sites, under the assumption that they evolve neutrally. We model a
355 population that splits at time T before present into two populations that exponentially
356 grow in size until the present day with sizes N_a , and N_b . We confirmed the timing
357 of the population splits between North Middle Atlas (representing Morocco) and
358 Western Europe and Central Asia separately. We used the same generation time and
359 mutation rate as in the MSMC analysis (1 year/generation; 7.1×10^{-9} mutations per
360 base pair per generation). We estimated uncertainty using the Godambe Information
361 Matrix (35). These results are reported in Table S5.

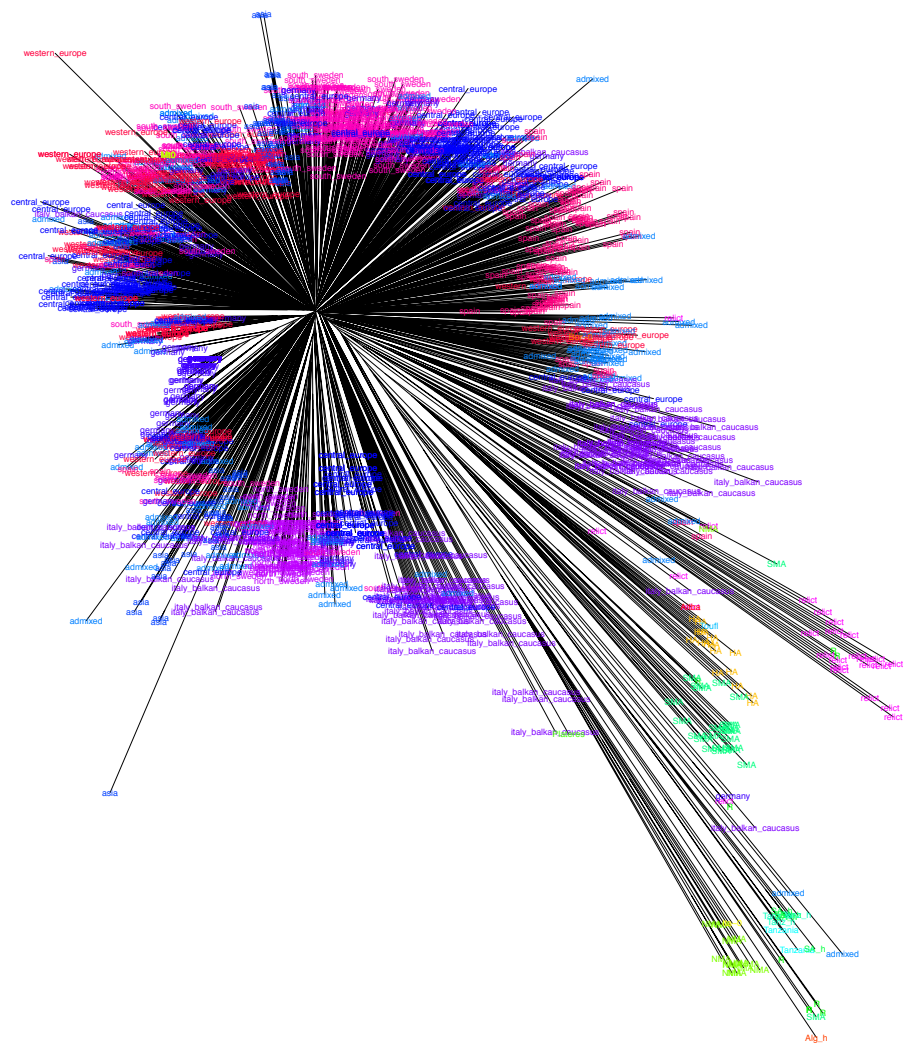
362
363
364

365

366 **SUPPLEMENTARY FIGURES**

367

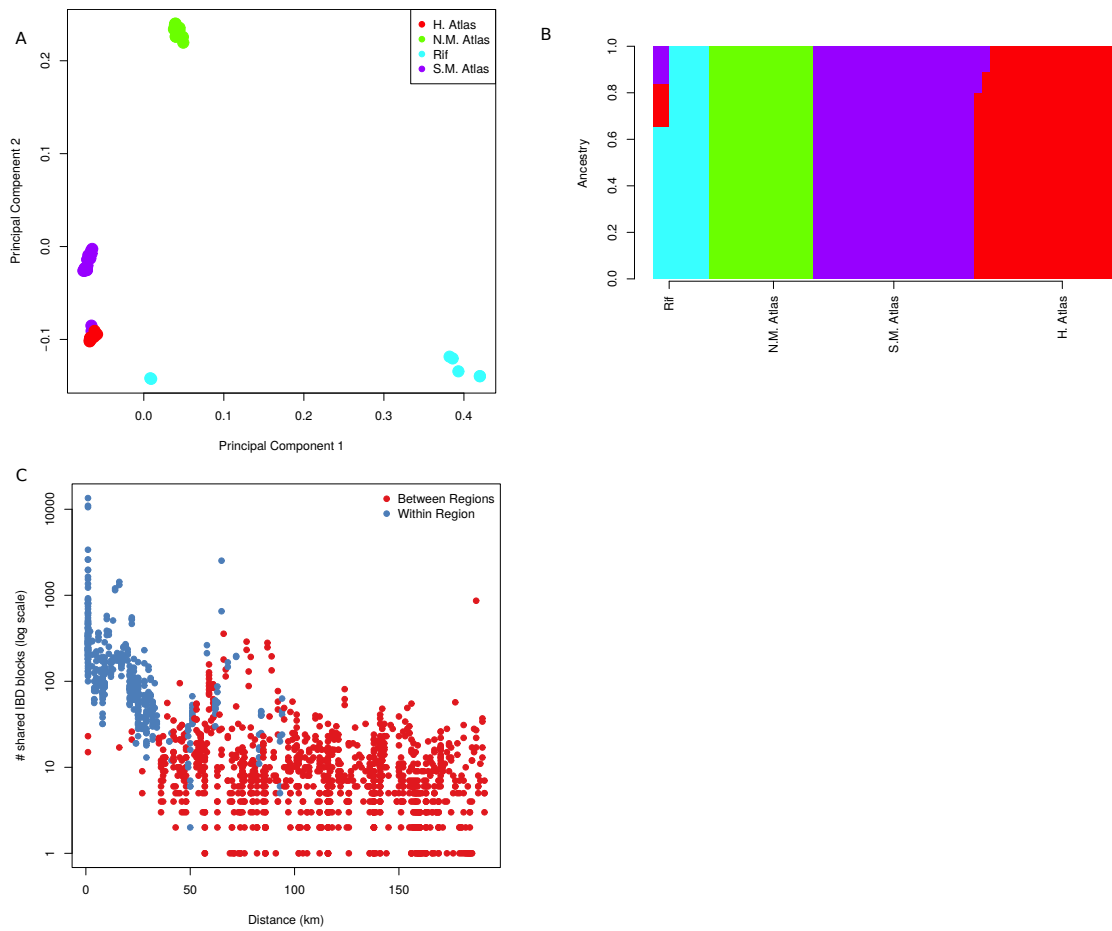
368



369

370 **Figure S1. Neighbor-Joining tree with all samples, colored by cluster.**

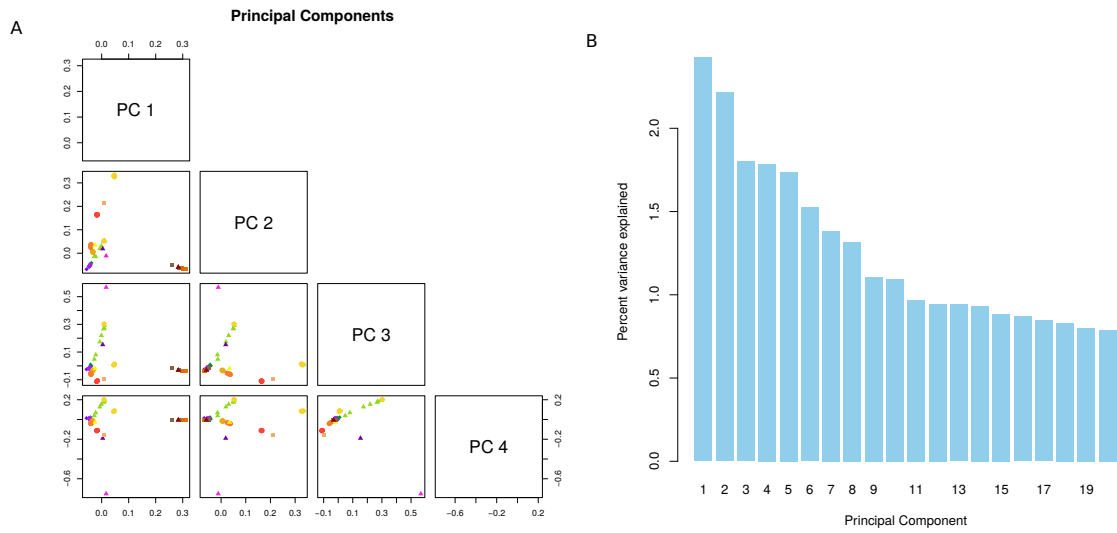
371



372
 373
 374
 375
 376
 377
 378
 379
 380
 381

Figure S2. Population structure in Morocco.

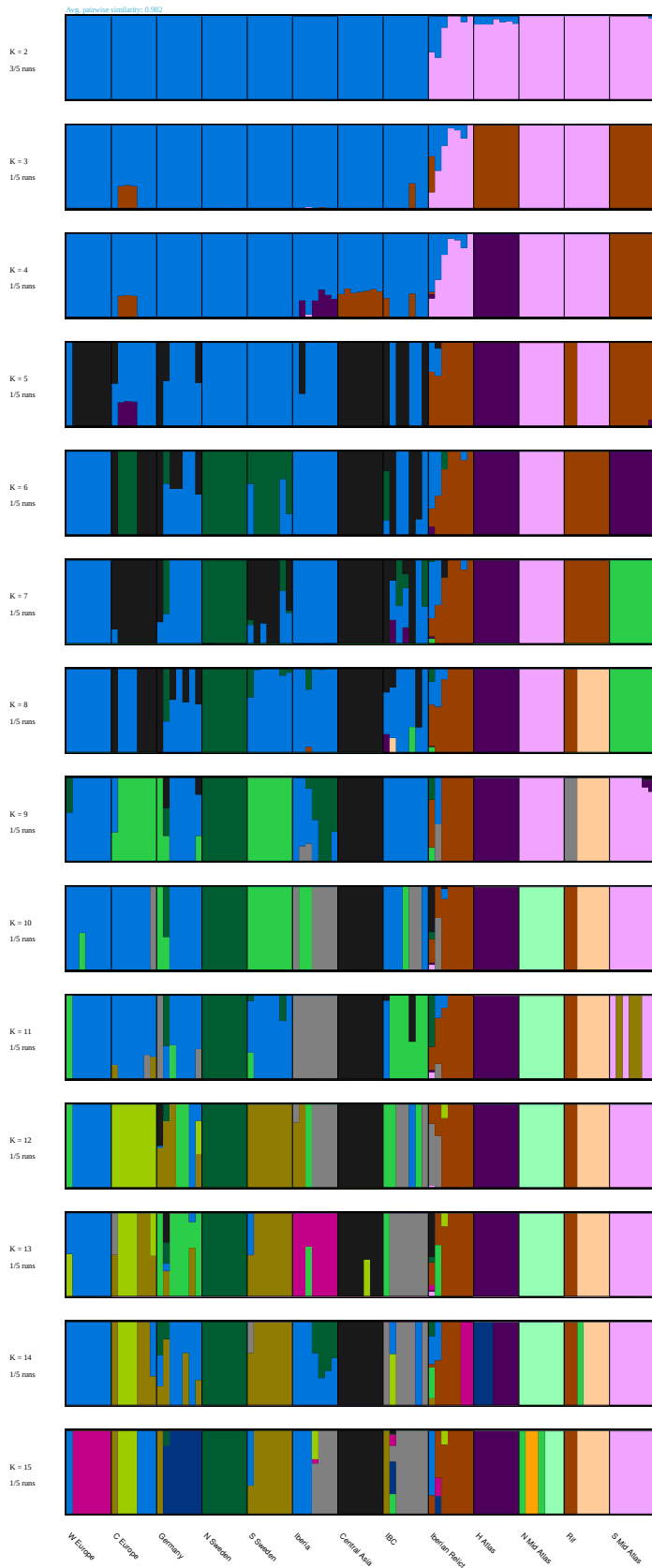
A) Principal components analysis shows strong population structure in Morocco. B) ADMIXTURE with the lowest cross-validation error ($K=4$) confirms the strong population structure. C) Geographic decay of shared haplotypes calculated using RefinedIBD in Beagle.



382
 383
 384
 385
 386
 387
 388

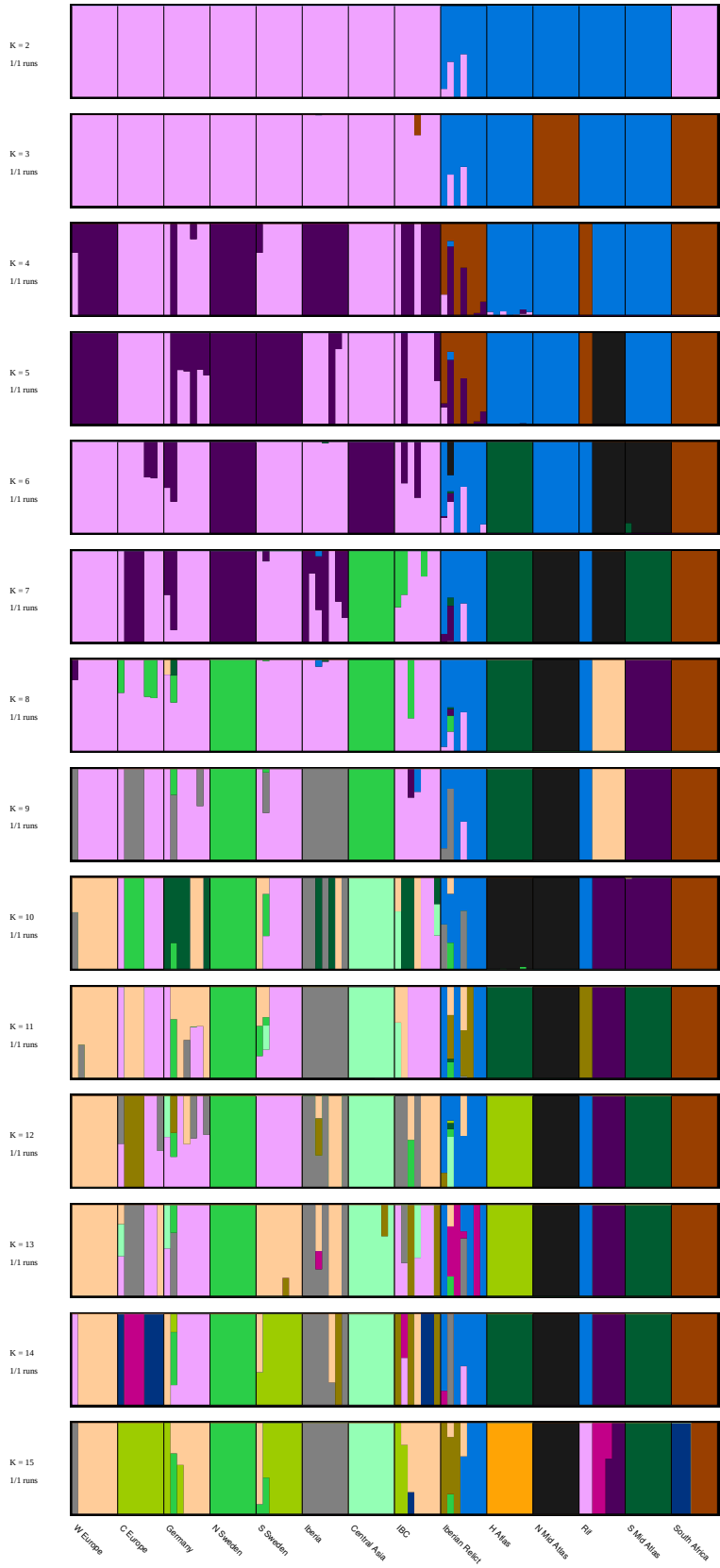
Figure S3. Worldwide PCA.

A) Worldwide PCA for the first 4 PCs. B) Proportion of variance explained by each PC up to PC 20.



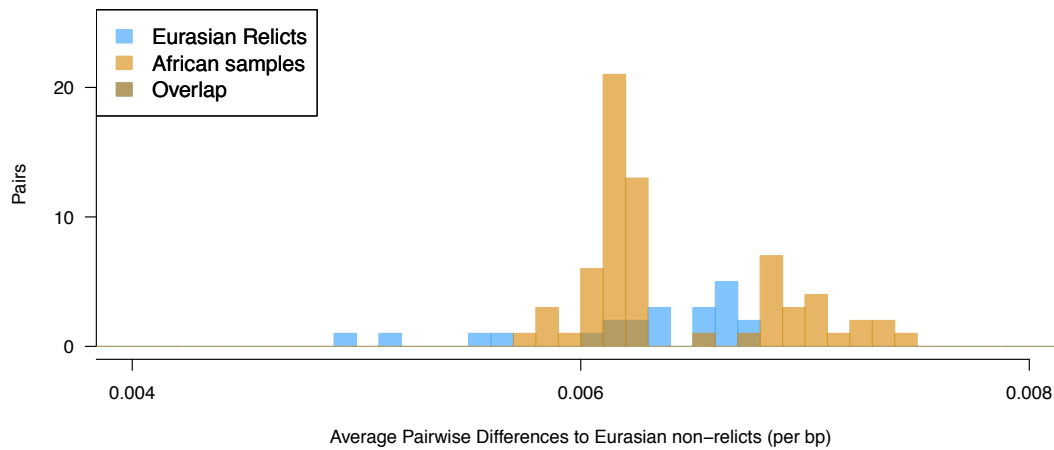
389
390
391
392

Figure S4. Worldwide ADMIXTURE for K=2 to K=15.
The run with the lowest cross-validation error (out of 20 replicates) is plotted.



393
 394
 395
 396

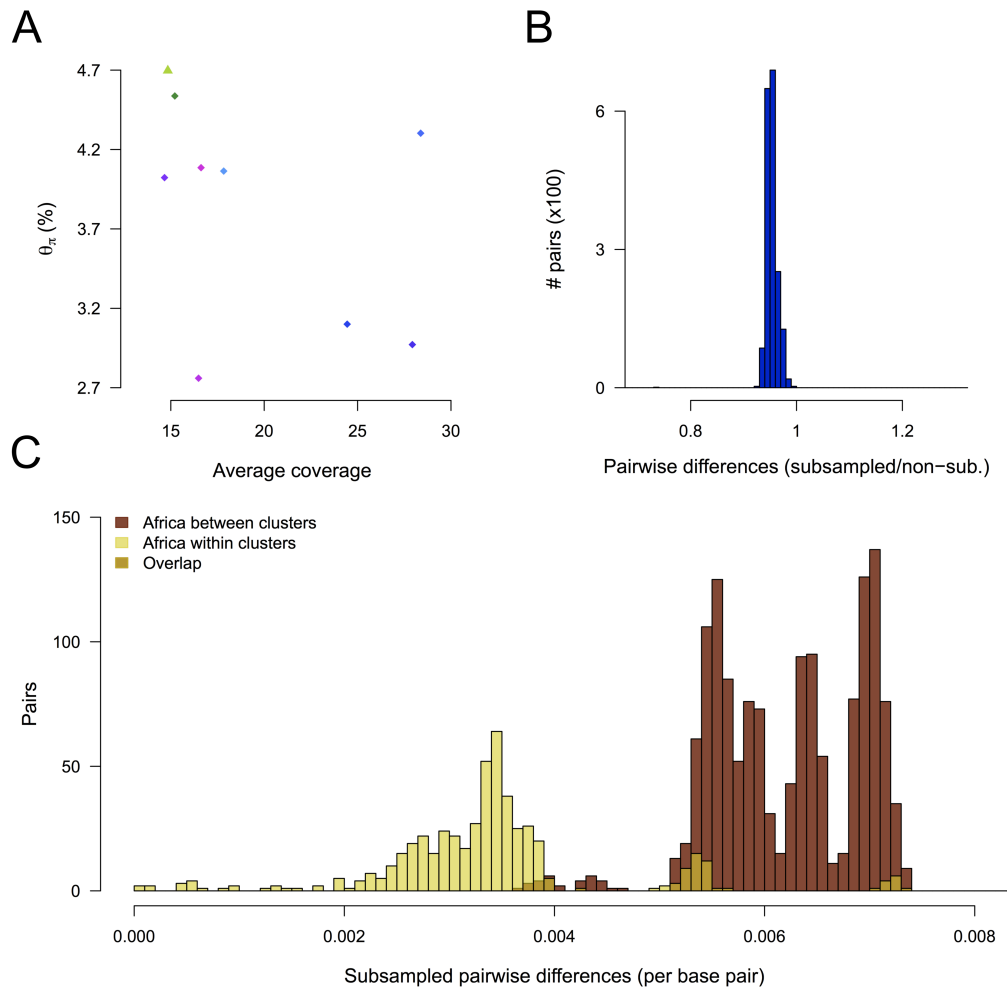
Figure S5. Worldwide ADMIXTURE including sub-Saharan samples for K=2 to K=15.



397
 398
 399
 400
 401
 402
 403
 404
 405

Figure S6. Divergence from Eurasian non-relicts.

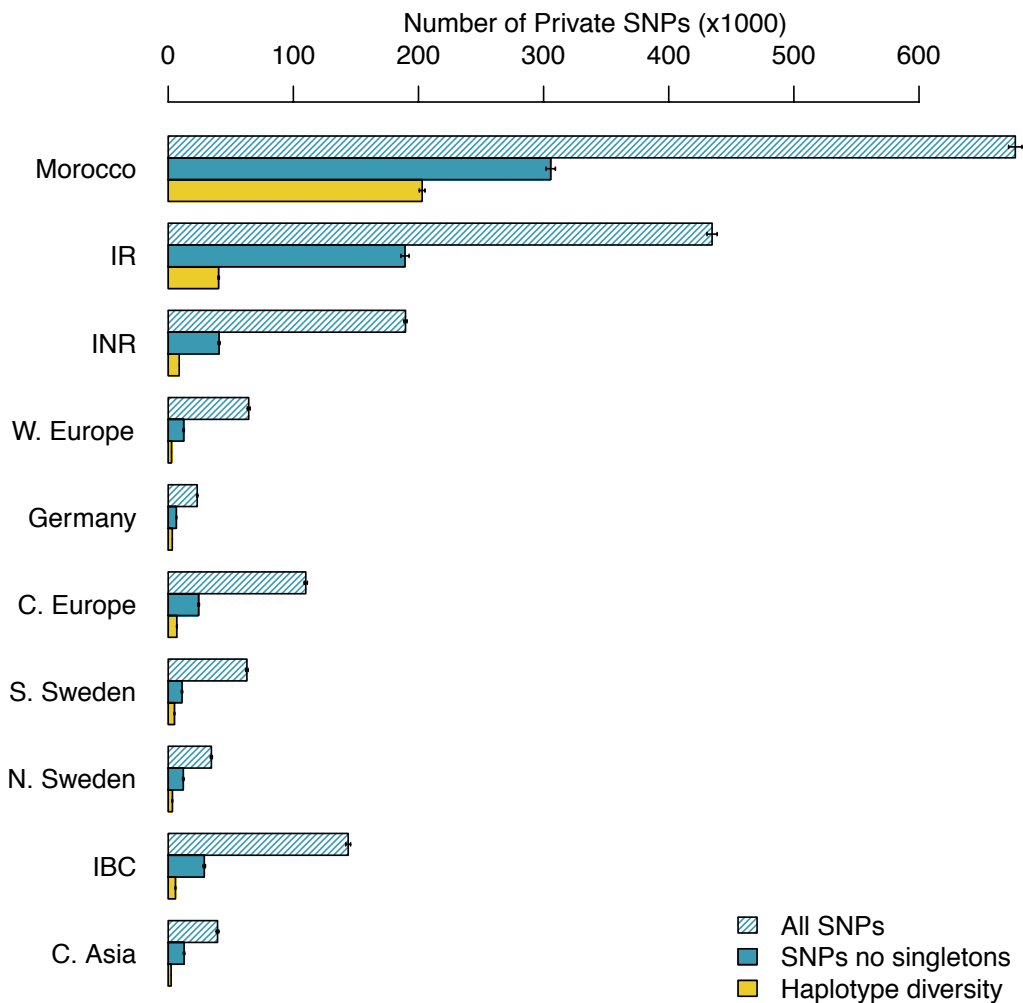
Average pairwise differences was used as a measure of divergence of Eurasian relicts (blue) and Africans (tan) relative to Eurasian non-relicts. Every African accession meets the criterion for being defined as relict, and a subset of them are more diverged than any Eurasian relict.



406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416
 417

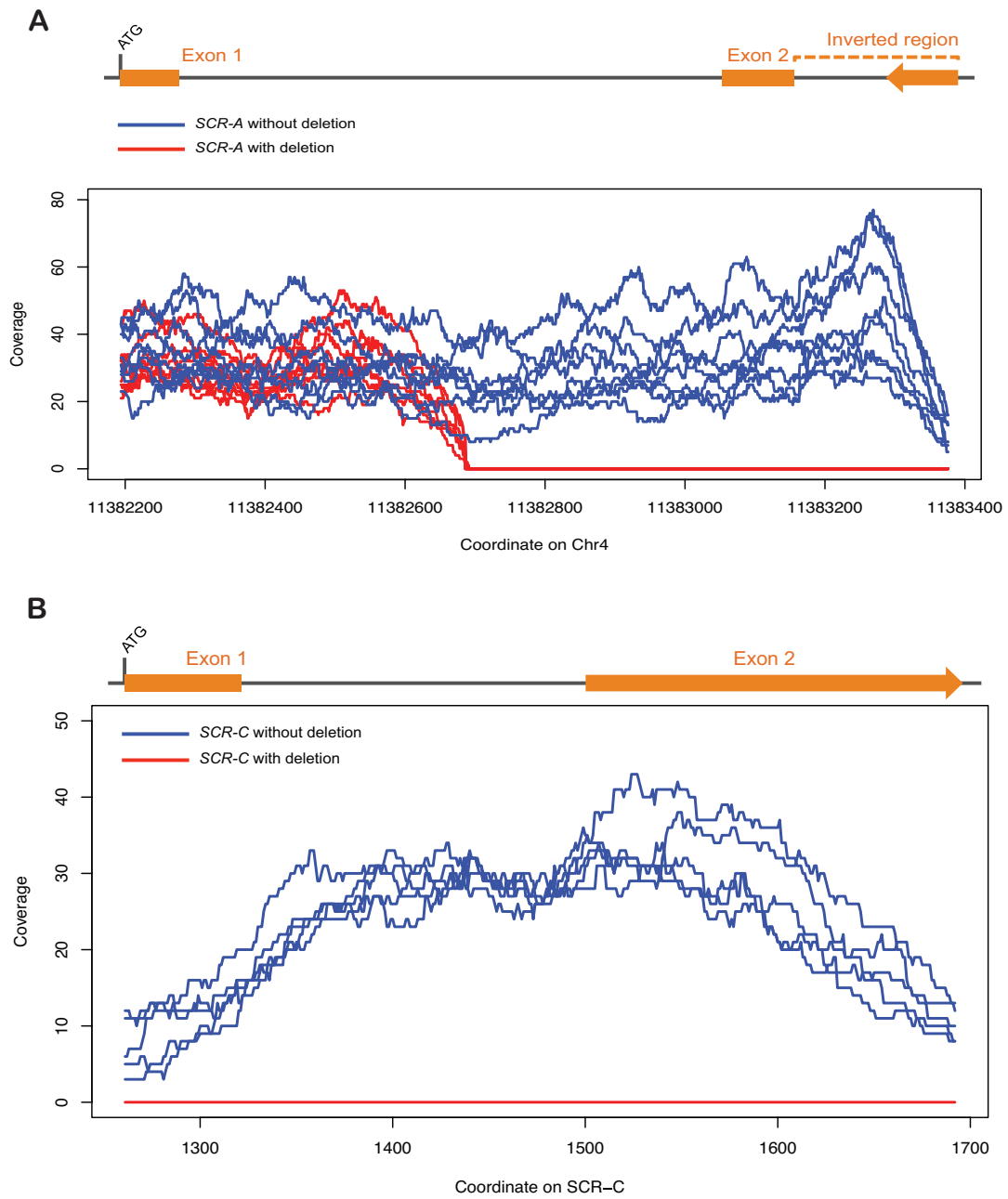
Figure S7. Effect of average coverage on diversity.

a) Diversity within clusters (average pairwise differences, θ_{π}) as a function of average coverage for all Eurasian clusters. There is no significant correlation between these quantities. b) Ratio between pairwise differences among African samples after artificially reducing coverage to 15x, the minimum average coverage among Eurasian clusters (subsampled), and among full-coverage samples (non-sub.). c) Distribution of pairwise differences per base pair for all African samples after reducing coverage to 15x (cfr. Fig. 2B for full-depth comparisons).



418
 419
 420
 421
 422
 423
 424
 425
 426
 427

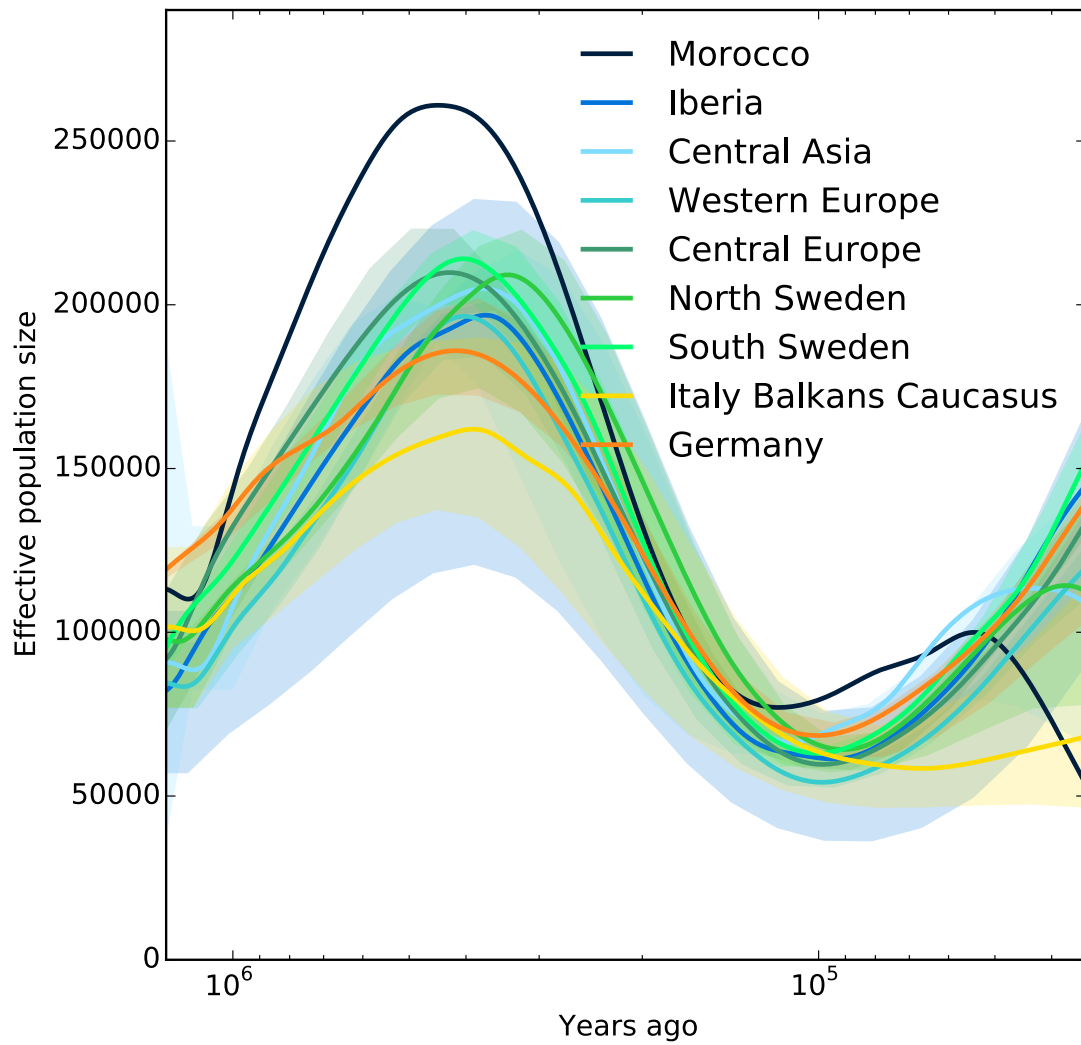
Figure S8. Private variation with Morocco considered as a single clade. Numbers of private SNPs with singletons (hashed blue) and without singletons (dark blue), and number of SNPs without singletons on private haplotypes (yellow) within each cluster. Error bars show 95% confidence intervals from calculations across 500 resampled datasets. Morocco harbors the highest diversity in all measures of private variation.



428
429

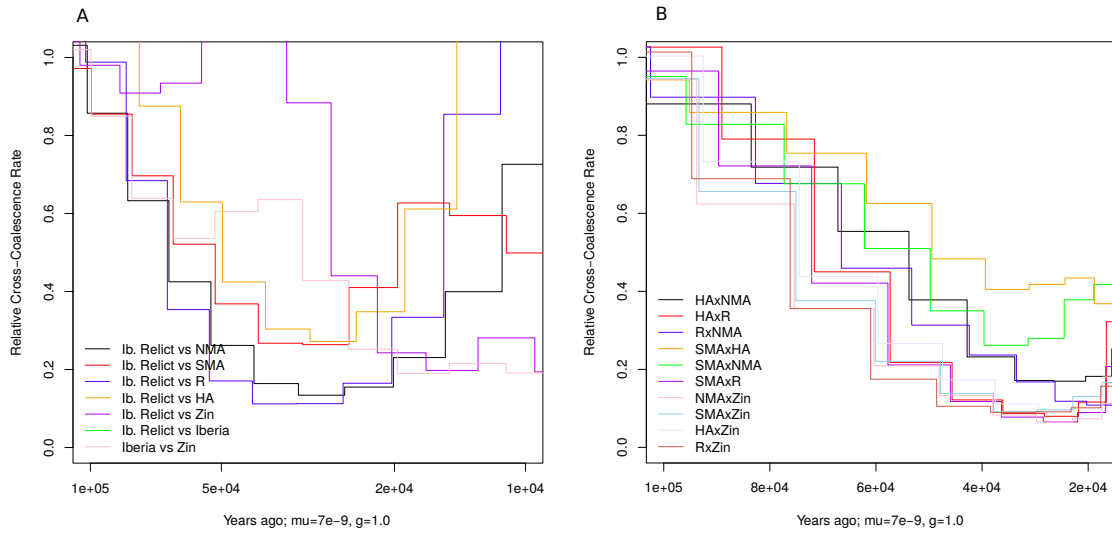
430 **Figure S9. Novel deletion haplotypes for S-locus haplogroups A and C in**
431 **Morocco.**

432 A) Schematic of the novel deletion haplotype for S-locus haplogroup A (coverage
433 drops to zero at the deletion). B) Novel deletion haplotype for S- locus haplogroup C
434 interesting the region of the SCR gene.
435



436
 437
 438
 439
 440

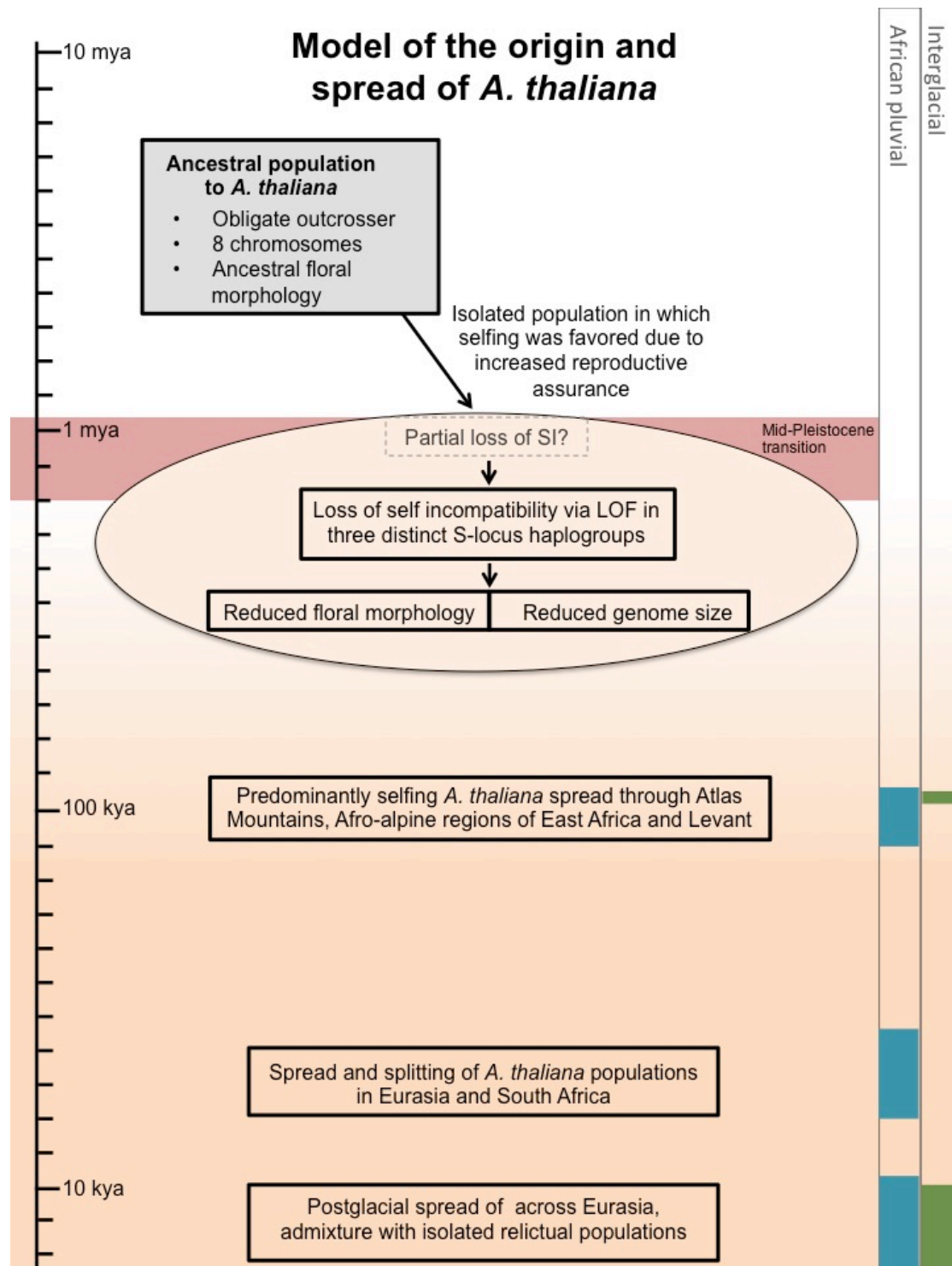
Figure S10. MSMC for Morocco and all Eurasian non-relict clusters.
 Effective population size medians are plotted with ± 1 standard deviation shaded.



441
 442
 443
 444
 445
 446
 447
 448
 449

Figure S11. Relative cross coalescence for comparisons between regions within Morocco and Iberian Relicts.

A) Relative cross-coalescence for comparisons between Iberian relicts and Individual Moroccan regions. The Rif-Zin population shows the most recent population continuity with the Iberian relicts, consistent with results from PCA and ADMIXTURE. B) Cross coalescence results for comparisons within Morocco.



450
451

452 **Figure S12.** A graphical model of the origin of selfing and spread of *A. thaliana*.
453 Inferred events are shown in boxes. Red bar denotes the Middle Pleistocene
454 Transition (~1.2-0.8 mya) and pluvials and interglacials over the past 100 kya are
455 shown in bars on the right.

456

457 The proposed model is as follows:

458

458 First, a sub-population became geographically isolated from the ancestral outcrossing
459 *A. thaliana* population. Based on our data, the most plausible scenario is that this
460 separation was due to migration of this subpopulation into Africa by 1.2-0.8 mya.

461 This timing corresponds to the Middle Pleistocene Transition, a shift to more arid
462 climates, more open habitats in Africa (woodlands to grasslands), and the beginning
463 of more severe glacial cycles worldwide (36, 37). This also corresponds to previous
464 estimates for the timing of the transition to selfing (29, 38) on the most current
465 mutation rate estimate (17, 39). This is similar to the situation suggested in *A. lyrata*
466 (40), where partial selfing arose multiple times in the same geographic region in
467 North America and may have been aided by the bottleneck itself. The rationale for
468 this is that during a strong bottleneck, inbreeding would have increased and genetic
469 load, a major impediment to the evolution of selfing in general, would have been
470 purged (41). In addition, the reduction in S-locus variation that likely occurred as a
471 result of the bottleneck would have reduced the probability of compatible matings in
472 the founding population, which would have favored selfing.

473
474 Then, we infer that the three distinct S-locus loss of function events occurred in this
475 isolated population, best represented today within Morocco. Partial loss of self-
476 incompatibility may have preceded the transition to predominant selfing in *A. thaliana*
477 via genetic changes outside of the S-locus (e.g., at ARC1 (42)).

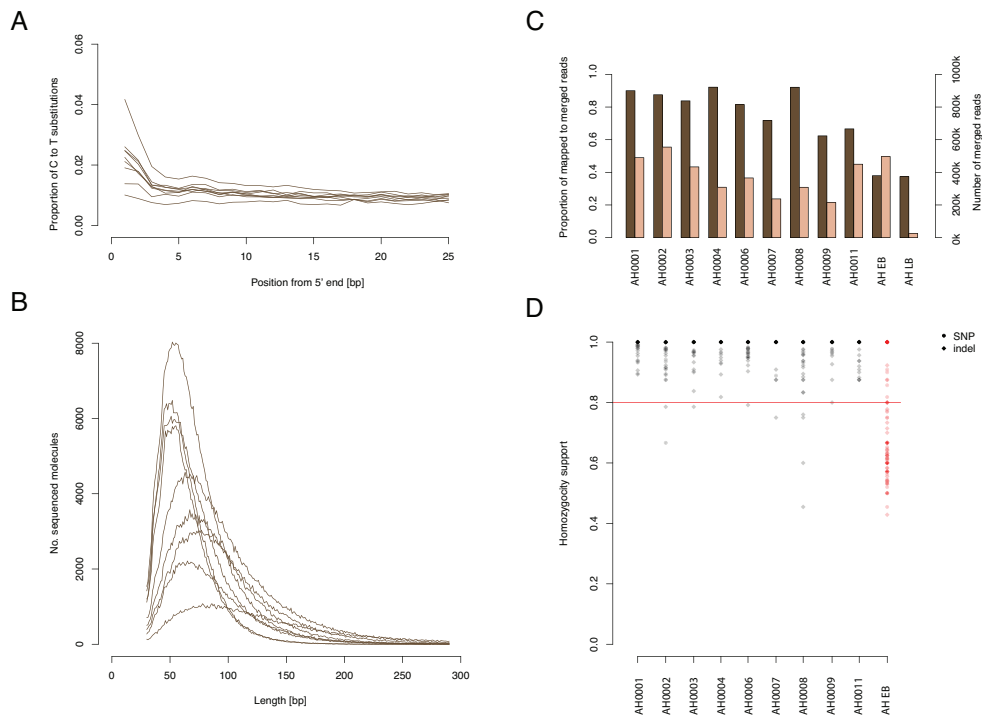
478
479 Next, subsequent to the transition to predominant selfing, the severely reduced floral
480 morphology found in *A. thaliana* evolved (43). In addition, reduced genome size
481 likely occurred at this point (44, 45), as has been observed in other cases after the
482 transition to predominant selfing (46).

483
484 Around 120 kya, during the Abbassia Pluvial and last interglacial, populations
485 expanded throughout Africa (Morocco and East African mountains) and into the
486 Levant. This would have happened at a time when there was a general increase in
487 vegetation levels across Africa (47) as well as in Eurasia (48, 49). It is also around the
488 same time that modern humans are thought to have first migrated out of Africa (50,
489 51).

490
491 Following this, at around 45-35 kya, we find evidence of splitting between
492 populations from Western Europe and Asia, which is similar to the timing previously
493 observed for analogous migrations in human populations (52-54).

494
495 Finally, after the end of the last glacial period, there is evidence of postglacial spread
496 across Eurasia and admixture with relict populations (27, 55-58).

497



1
2
3
4
5
6
7
8
9
10
11

Figure S13. Authentication and analysis of herbarium ancient DNA (aDNA) samples.

A) Proportion of C-T transitions at the 5'-end of the reads. B) Read length distribution. C) Fraction of *A. thaliana* DNA as a proportion of mapped reads to merged reads (brown bars), and total number of merged reads (beige bars) per sample. AHEB and AHLB denote controls. AHEB denotes extraction blank and AHLB denotes library blank. D) Homozygosity support for SNPs and indels called in chloroplast genome.

12 **SUPPLEMENTARY TABLES**
13

Table S1. Sequenced samples. Samples sequenced from fresh material and herbarium material are listed together with sample location information and average coverage. ‘Country’ denotes two-letter ISO Country Codes. ‘Cov¹’ shows Golden Path coverage and ‘Cov²’ shows coverage of non-missing bases. Ket-10 and Ma-0 were sequenced multiple times and the comparison was used for error rate estimates.

Sample ID	Country	Region	Cov ¹	Cov ²	Latitude	Longitude	Herbarium - collection year
Ait14	MA	HA	35.22	40.95	31.2366	-7.81285	-
Ait9	MA	HA	20.88	24.35	31.2366	-7.81285	-
Arb0	MA	HA	31.24	36.20	31.41988	-7.5262	-
Arb2	MA	HA	28.17	32.87	31.41988	-7.5262	-
Elh10	MA	HA	28.57	33.20	31.47197	-7.40644	-
Elh15	MA	HA	25.41	29.64	31.47197	-7.40644	-
Elh20	MA	HA	24.10	28.07	31.47197	-7.40644	-
Elh23	MA	HA	32.01	36.92	31.47197	-7.40644	-
Elh2	MA	HA	27.08	31.63	31.47197	-7.40644	-
Elh27	MA	HA	22.93	26.50	31.47197	-7.40644	-
Elh33	MA	HA	30.04	34.85	31.47197	-7.40644	-
Elh39	MA	HA	26.52	30.96	31.47197	-7.40644	-
Elh46	MA	HA	37.24	43.19	31.47197	-7.40644	-
Elh52	MA	HA	27.80	32.30	31.47197	-7.40644	-
Set0	MA	HA	33.86	39.03	31.22656	-7.67361	-
Set6	MA	HA	28.76	33.47	31.22656	-7.67361	-
Bba0	MA	NMA	18.63	22.09	35.04256	-5.02369	-
Bba1	MA	NMA	24.80	29.38	34.01934	-4.0839	-
Bba2	MA	NMA	25.79	30.46	34.01934	-4.0839	-
Meh0	MA	NMA	25.86	30.30	33.9561	-4.05153	-
Meh4	MA	NMA	22.98	26.91	33.9561	-4.05153	-
Meh7	MA	NMA	23.22	27.43	33.9561	-4.05153	-
Tah0	MA	NMA	27.68	32.49	34.05293	-4.22245	-
Tah4	MA	NMA	25.69	30.28	34.05293	-4.22245	-
Taz0	MA	NMA	22.66	26.72	34.09166	-4.10258	-
Taz11	MA	NMA	31.06	36.59	34.09166	-4.10258	-
Taz16	MA	NMA	23.01	30.67	34.09166	-4.10258	-
Taz18	MA	NMA	28.21	33.27	34.09166	-4.10258	-
Tiz0	MA	NMA	26.75	31.45	33.8723	-4.02647	-
Tiz7	MA	NMA	28.44	33.54	33.8723	-4.02647	-
Bab0	MA	Rif	21.78	25.68	35.04256	-5.02369	-
Bab3	MA	Rif	32.95	38.84	35.04256	-5.02369	-
Bbe0	MA	Rif	32.91	38.66	34.99519	-4.83141	-
Ket10 (1)	MA	Rif	23.16	27.35	34.96076	-4.66611	-
Ket10 (2)	MA	Rif	27.79	32.88	34.96076	-4.66611	-
Ket12	MA	Rif	15.81	18.93	34.96076	-4.66611	-
Zin4	MA	Rif	24.71	28.99	35.4528	-5.42698	-
Zin9	MA	Rif	22.37	26.16	35.4528	-5.42698	-
Agl0	MA	SMA	25.22	29.51	32.97243	-5.44856	-

Agl1	MA	SMA	28.35	33.21	32.97243	-5.44856	-
Agl2	MA	SMA	25.29	29.71	32.97243	-5.44856	-
Agl3	MA	SMA	26.47	31.02	32.97243	-5.44856	-
Agl5	MA	SMA	29.47	34.55	32.97243	-5.44856	-
Agl9	MA	SMA	26.49	31.09	32.97243	-5.44856	-
Azr0	MA	SMA	20.52	24.00	33.42357	-5.17911	-
Azr11	MA	SMA	28.07	32.73	33.42357	-5.17911	-
Azr13	MA	SMA	28.08	32.97	33.42357	-5.17911	-
Azr16	MA	SMA	24.69	28.89	33.42357	-5.17911	-
Azr5	MA	SMA	26.76	31.22	33.42357	-5.17911	-
Azr7	MA	SMA	29.05	33.92	33.42357	-5.17911	-
Elk1	MA	SMA	29.20	34.13	32.53516	-6.014969	-
Elk20	MA	SMA	15.12	17.99	32.53516	-6.014969	-
Elk28	MA	SMA	30.51	35.78	32.53516	-6.014969	-
Elk3	MA	SMA	15.43	18.31	32.53516	-6.014969	-
Elk5	MA	SMA	19.74	23.30	32.53516	-6.014969	-
IFr0	MA	SMA	27.43	32.01	33.55006	-5.17465	-
Ifr3	MA	SMA	28.39	33.25	33.55006	-5.17465	-
Ifr4	MA	SMA	25.07	29.52	33.55006	-5.17465	-
Ifr6	MA	SMA	29.23	34.36	33.55006	-5.17465	-
Khe0	MA	SMA	24.42	28.54	32.92735	-5.51027	-
Khe32	MA	SMA	26.39	30.93	32.92735	-5.51027	-
Oua0	MA	SMA	23.23	27.05	32.07853	-6.275309	-
Til2	MA	SMA	34.70	40.53	32.04208	-6.22955	-
Toufl-1	MA	MA	24.36	28.30	31.47	-7.42	-
Aitba	MA	MA	37.46	43.14	31.48	-7.45	-
Ita-0	MA	MA	22.86	27.21	34.0787	-4.19891	-
Tanz-1	TZ	TZ	25.92	30.93	-2.8739	36.12	-
Tanz-2	TZ	TZ	34.78	41.21	-2.8739	36.12	-
SA-h1	ZA	ZA	7.76	10.43	-34.125	19.375	1830
SA-h2	ZA	ZA	9.04	11.83	-34.125	19.375	1830
SA-h3	ZA	ZA	8.28	10.96	-34.125	19.375	1830
SA-h4	ZA	ZA	7.28	9.70	33.399	19.282	1896
SA-h5	ZA	ZA	6.70	9.24	-34.125	19.375	1830
Tanz-h1	TZ	TZ- Kilimanjaro	8.90	11.78	-3.021568	37.159565	1929
Tanz-h2	TZ	TZ-Mt. Meru	9.09	12.07	-3.23333	36.71667	1985
Tanz-h3	TZ	TZ-Mt. Meru	5.41	7.89	-3.23333	36.71667	1985
Alg-h1	DZ	DZ	4.54	6.78	-	-	1837
Platres	CY	CY	26.21	30.31	34.8833	32.8666	-
Ma0 (1)	DE	DE	36.90	41.64	50.8167	8.7667	-
Ma0 (2)	DE	DE	42.91	48.32	50.8167	8.7667	-
Ma0 (3)	DE	DE	31.79	35.96	50.8167	8.7667	-
Ma0 (4)	DE	DE	41.32	46.57	50.8167	8.7667	-

Table S2. Average cross-validation error across ADMIXTURE runs and lowest CV error for each K.

K	Mean CV Error (n=20)	Min CV Error
2	0.0988669	0.09846
3	0.0977854	0.09596
4	0.09751	0.0952
5	0.0982888	0.09459
6	0.0988346	0.09574
7	0.0960972	0.09662
8	0.0975108	0.09825
9	0.103088	0.09976
10	0.104702	0.10145
11	0.171991	0.10428
12	0.110581	0.10687
13	0.112384	0.10784
14	0.116485	0.11161
15	0.119311	0.11534

Table S3. Significance tests for the number of private variants per cluster. We tested whether Moroccan and Eurasian clusters had significantly different numbers of private variants using the Welch's unequal variance, two sample *t*-test. The four columns referring to separate Moroccan clusters are relevant to the results displayed in Fig. 2C. The column for Morocco as a whole is relevant to the results shown in Fig. S6. For each of the three measures of private variation, the *t*-statistic (*t*), degrees of freedom (*d.f.*) and *p*-value (*p*) of the analysis are shown.

		Rif	NMA	SMA	HA	MOR	
All SNPs	Central Asia	<i>t</i>	108.55	150.32	159.03	114.56	229.30
		<i>d.f.</i>	527.82	628.09	642.70	707.61	534.56
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Iberia, Balkans, Caucasus	<i>t</i>	88.42	101.52	106.38	59.19	183.78
		<i>d.f.</i>	601.45	884.43	911.10	983.93	624.11
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Northern Sweden	<i>t</i>	109.13	150.92	159.56	115.36	233.09
		<i>d.f.</i>	530.79	640.88	656.80	726.87	516.73
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Southern Sweden	<i>t</i>	105.41	143.79	152.15	106.46	221.38
		<i>d.f.</i>	528.59	631.41	646.36	712.64	529.71
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Central Europe	<i>t</i>	95.95	121.03	127.61	79.61	202.44
		<i>d.f.</i>	553.19	732.10	755.87	850.95	549.82
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Germany	<i>t</i>	112.62	163.31	173.56	130.78	238.44
		<i>d.f.</i>	506.78	534.39	538.55	557.78	506.21
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Western Europe	<i>t</i>	104.85	141.34	149.34	103.66	219.73
		<i>d.f.</i>	535.87	662.44	680.46	758.41	540.33
		<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
Iberian non-relicts	<i>t</i>	83.30	99.55	105.85	50.99	174.02	
	<i>d.f.</i>	522.68	605.60	617.82	672.90	549.51	
	<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	
Iberian relicts	<i>t</i>	19.63	-10.38	-9.92	-43.30	70.97	
	<i>d.f.</i>	960.68	794.19	768.27	688.03	925.26	
	<i>p</i>	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	

			Rif	NMA	SMA	HA	MOR
SNPs no singletons	Central Asia	t	65.28	125.24	81.69	68.50	151.91
		d.f.	501.00	513.32	512.33	526.23	516.35
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Iberia, Balkans, Caucasus	t	64.76	122.50	79.64	65.36	140.79
		d.f.	504.89	541.18	538.28	578.90	556.07
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Northern Sweden	t	65.31	124.92	81.55	68.23	152.22
		d.f.	502.10	521.23	519.70	541.25	516.49
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Southern Sweden	t	65.90	127.01	83.34	70.87	153.12
		d.f.	500.62	510.60	509.80	521.08	511.74
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Central Europe	t	64.63	121.36	78.91	64.22	145.14
		d.f.	507.33	558.52	554.44	611.22	527.00
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Germany	t	65.47	125.95	82.28	69.39	156.16
		d.f.	500.45	509.40	508.68	518.79	502.76
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Western Europe	t	65.62	125.88	82.42	69.48	152.27
		d.f.	501.75	518.72	517.36	536.48	513.60
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
Iberian non-relicts	t	63.31	117.11	75.17	58.94	135.83	
	d.f.	509.85	576.33	571.06	643.87	539.99	
	p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	
Iberian relicts	t	44.04	47.21	24.95	6.95	45.76	
	d.f.	805.28	847.19	865.01	702.91	981.55	
	p	<2.2E-16	<2.2E-16	<2.2E-16	8.35E-12	<2.2E-16	

			Rif	NMA	SMA	HA	MOR
Haplotype Variation	Central Asia	t	70.97	161.13	138.18	259.49	174.34
		d.f.	499.23	501.35	505.54	604.36	499.34
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Iberia, Balkans, Caucasus	t	69.60	156.35	129.94	209.01	171.28
		d.f.	499.76	506.91	520.99	816.50	499.86
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Northern Sweden	t	70.55	159.09	134.24	211.09	173.36
		d.f.	500.21	511.61	534.04	929.38	500.49
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Southern Sweden	t	69.87	157.08	131.06	207.43	171.96
		d.f.	499.95	508.88	526.46	871.67	500.14
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Central Europe	t	69.15	154.02	125.50	167.22	170.22
		d.f.	501.05	520.41	558.37	997.81	501.52
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Germany	t	70.55	159.51	135.27	232.73	173.49
		d.f.	499.61	505.37	516.70	765.25	499.77
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Western Europe	t	70.83	160.61	137.25	250.96	174.05
		d.f.	499.33	502.48	508.69	653.05	499.37
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
Iberian non-relicts	t	68.39	152.73	124.14	201.35	168.71	
	d.f.	499.34	502.55	508.89	655.99	499.50	
	p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	
Iberian relicts	t	55.26	92.95	39.01	-7.74	137.19	
	d.f.	549.62	912.45	962.26	540.39	560.70	
	p	<2.2E-16	<2.2E-16	<2.2E-16	4.97E-14	<2.2E-16	

Table S4. Frequency of S-locus haplogroups among clusters. The frequencies of S-locus haplogroups in Morocco are very different from Eurasian clusters. S-locus haplogroup B occurs only in the Moroccan region Rif in the worldwide sample. Haplogroup C, rare in Eurasia, is very prevalent in two Moroccan regions. Its frequency is similar in Moroccan Rif and Iberian Relict populations, consistent with a Moroccan origin of Iberian Relicts. The recombinant haplogroup (Hap-AC), which is at high frequency in Eurasia, occurs at low frequency in Moroccan clusters.

Region	S-locus frequency (%)			
	Hap-A	Hap-AC	Hap-C	Hap-B
Rif	14.3	0.0	14.3	71.4
North Mid Atlas	50.0	14.3	35.7	0.0
South Mid Atlas	32.0	8.0	60.0	0.0
High Atlas	100.0	0.0	0.0	0.0
Western Europe	31.3	61.3	7.5	0.0
Italy, Balkans, Caucasus	61.5	32.3	6.2	0.0
Central Europe	70.0	22.5	7.5	0.0
Central Asia	94.1	0.0	5.9	0.0
Southern Sweden	73.2	21.8	4.9	0.0
Northern Sweden	75.0	25.0	0.0	0.0
Germany	29.8	68.9	1.2	0.0
Iberian Relict	77.3	4.5	18.2	0.0
Iberian non-relicts	50.7	45.1	4.2	0.0

Table S5. *δaδi* estimated parameters. We fit a simple isolation model in which two populations split at time T and exponentially grow to size N_a and N_b . We fit this model to the folded joint allele frequency spectra between North Middle Atlas (NMA) and Western Europe (WE) and NMA and Central Asia (CA).

	N_a	N_b	T
NMA-WE	17013.74 ± 2197.37	34714.41 ± 4527	128895.60 ± 12836.64
NMA-CA	19379.51 ± 994.30	33070.17 ± 1769.36	134847.93 ± 5744.91

Table S6. Error rate estimation for the two pipelines.

We independently sequenced the same Moroccan accession twice (Ket10), plus four replicates of accession Ma-0 for a total of seven pairs of putatively identical accessions. Results from the two pipelines are shown, in terms of overlap of the genome called in both replicates, and differences between replicates. FulgiPipe has a much lower error rate, at the cost of disregarding a larger proportion of the genome.

	Pairs	Overlap non-missing (bp)	# differences	Error rate (bp ⁻¹)
FulgiPype	Ket10 ₁ -Ket10 ₂	89402238	15	1.68E-07
	Ma0 ₁ -Ma0 ₂	77690227	6	7.72E-08
	Ma0 ₁ -Ma0 ₃	72829851	7	9.61E-08
	Ma0 ₁ -Ma0 ₄	77908083	10	1.28E-07
	Ma0 ₂ -Ma0 ₃	76179368	6	7.88E-08
	Ma0 ₂ -Ma0 ₄	88134953	8	9.08E-08
	Ma0 ₃ -Ma0 ₄	76414793	4	5.23E-08
Average FP		79794216.1	8.0	9.88E-08
Shore-Mpi	Ket10 ₁ -Ket10 ₂	101319503	21491	2.12E-04
	Ma0 ₁ -Ma0 ₂	106836045	13082	1.22E-04
	Ma0 ₁ -Ma0 ₃	106501521	13242	1.24E-04
	Ma0 ₁ -Ma0 ₄	106790012	12976	1.22E-04
	Ma0 ₂ -Ma0 ₃	106626309	13174	1.24E-04
	Ma0 ₂ -Ma0 ₄	106952142	12838	1.20E-04
	Ma0 ₃ -Ma0 ₄	106584152	12958	1.22E-04
Average SM		105944240.6	14251.6	1.35E-04

References

1. Brennan AC, *et al.* (2014) The genetic structure of *Arabidopsis thaliana* in the south-western Mediterranean range reveals a shared history between North Africa and southern Europe. *BMC Plant Biol* 14:17.
2. Yoshida K, *et al.* (2013) The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *Elife* 2:e00731.
3. Briggs AW, *et al.* (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* 38(6):e87.
4. Meyer M & Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* 2010(6):pdb prot5448.
5. Meyer M, *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222-226.
6. Kircher M, Sawyer S, & Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40(1):e3.
7. Briggs AW, *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A* 104(37):14616-14621.
8. Jonsson H, Ginolhac A, Schubert M, Johnson PL, & Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29(13):1682-1684.
9. Weiß CL, *et al.* (2016) Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *Royal Society Open Science* 3(6).
10. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
11. Schubert M, Lindgreen S, & Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* 9:88.
12. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
13. Jiang H, Lei R, Ding SW, & Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15:182.
14. Magoc T & Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957-2963.
15. Ossowski S, *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18(12):2024-2033.
16. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
17. Ossowski S, *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327(5961):92-94.
18. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genet* 5(10):e1000686.
19. Puechmaille SJ (2016) The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol Ecol Resour* 16(3):608-627.

20. Shringarpure S & Xing EP (2014) Effects of sample selection bias on the accuracy of population structure and ancestry inference. *G3 (Bethesda)* 4(5):901-911.
21. Purcell S, *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559-575.
22. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24(11):1403-1405.
23. Jombart T & Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27(21):3070-3071.
24. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289-290.
25. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655-1664.
26. Browning BL & Browning SR (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194(2):459-471.
27. Consortium G (2016) 1135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166(2):481-491.
28. Karney CFF (2013) Algorithms for geodesics. *J. Geodesy* 87:43-55.
29. Tang C, *et al.* (2007) The evolution of selfing in *Arabidopsis thaliana*. *Science* 317(5841):1070-1072.
30. Dwyer KG, *et al.* (2013) Molecular characterization and evolution of self-incompatibility genes in *Arabidopsis thaliana*: the case of the Sc haplotype. *Genetics* 193(3):985-994.
31. Sherman-Broyles S, *et al.* (2007) S locus genes and the evolution of self-fertility in *Arabidopsis thaliana*. *Plant Cell* 19(1):94-106.
32. Schiffels S & Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46(8):919-925.
33. Bomblies K, *et al.* (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet* 6(3):e1000890.
34. Gutenkunst RN, Hernandez RD, Williamson SH, & Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.
35. Coffman AJ, Hsieh PH, Gravel S, & Gutenkunst RN (2016) Computationally Efficient Composite Likelihood Statistics for Demographic Inference. *Mol Biol Evol* 33(2):591-593.
36. Cerling TE & Hay RL (1986) An isotopic study of paleosol carbonates from Olduvai Gorge. *Quaternary Research* 25:63-78.
37. deMenocal PB (2004) African climate change and faunal evolution during the Pliocene-Pleistocene *Earth and Planetary Science Letters* 220:3-24.
38. Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, & Schierup MH (2006) The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. *Mol Biol Evol* 23(9):1741-1750.

39. Shimizu KK & Tsuchimatsu T (2015) Evolution of selfing: recurrent patterns in molecular adaptation. *Annual Review of Ecology Evolution and Systematics* 46:593-622.
40. Mable BK, *et al.* (2017) What causes mating system shifts in plants? *Arabidopsis lyrata* as a case study. *Heredity (Edinb)* 118(1):110.
41. Barrett SC & Charlesworth D (1991) Effects of a change in the level of inbreeding on the genetic load. *Nature* 352(6335):522-524.
42. Indriolo E, Safavian D, & Goring DR (2014) The ARC1 E3 Ligase Promotes Two Different Self-Pollen Avoidance Traits in *Arabidopsis*. *Plant Cell* 26(4):1525-1543.
43. Sicard A & Lenhard M (2011) The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Ann Bot* 107(9):1433-1443.
44. Hu TT, *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 43(5):476-481.
45. Oyama RK, *et al.* (2008) The shrunken genome of *Arabidopsis thaliana*. *Plant Systematics and Evolution* 273(3):257-271.
46. Wright SI, Ness RW, Foxe JP, & Barrett SCH (2008) Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences* 169(1):105-118.
47. Dupont L (2011) Orbital scale vegetation change in Africa. *Quaternary Science Reviews* 30(25-26):3589-3602.
48. Kaspar F, Kühl N, Cubasch U, & Litt T (2005) A model-data comparison of European temperatures in the Eemian interglacial. *Geophysical Research Letters* 32(11):n/a-n/a.
49. Kukla GJ, *et al.* (2002) Last Interglacial Climates. *Quaternary Research* 58:2-13.
50. Osborne AH, *et al.* (2008) A humid corridor across the Sahara for the migration of early modern humans out of Africa 120,000 years ago. *Proc Natl Acad Sci U S A* 105(43):16444-16447.
51. Timmermann A & Friedrich T (2016) Late Pleistocene climate drivers of early human migration. *Nature*.
52. Henn BM, Cavalli-Sforza LL, & Feldman MW (2012) The great human expansion. *Proc Natl Acad Sci U S A* 109(44):17758-17764.
53. Mallick S, *et al.* (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*.
54. Voight BF, *et al.* (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* 102(51):18508-18513.
55. Beck JB, Schmuths H, & Schaal BA (2008) Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Mol Ecol* 17(3):902-915.
56. Francois O, Blum MG, Jakobsson M, & Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet* 4(5):e1000075.
57. Pico FX, Mendez-Vigo B, Martinez-Zapater JM, & Alonso-Blanco C (2008) Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula. *Genetics* 180(2):1009-1021.

58. Sharbel TF, Haubold B, & Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol Ecol* 9(12):2109-2118.