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

Arun Durvasula[a,b,c,1], Andrea Fulgione[a,b,c,1], Rafal M. Gutaker[d], Selen Irez Alacakaptan[b,c], Pádraic J. Flood[a], Célia Neto[a], Takashi Tsuchimatsu[e], Hernán A. Burbano[d], F. Xavier Picó[f], Carlos Alonso-Blanco[g], and Angela M. Hancock[a,b,c,2]

[a]Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany; [b]Department of Structural and Computational Biology, University of Vienna, 1010 Vienna, Austria; [c]Vienna Biocenter, 1030 Vienna, Austria; [d]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; [e]Department of Biology, Chiba University, Chiba 263-8522 Japan; [f]Departamento de Ecología Integrativa, Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, 41092 Seville, Spain; and [g]Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain

Edited by Johanna Schmitt, University of California Davis, CA, and approved April 11, 2017 (received for review October 13, 2016)

Over the past 20 y, many studies have examined the history of the plant ecological and molecular model, *Arabidopsis thaliana*, in Europe and North America. Although these studies informed us about the recent history of the species, the early history has remained elusive. In a large-scale genomic analysis of African *A. thaliana*, we sequenced the genomes of 78 modern and herbarium samples from Africa and analyzed these together with over 1,000 previously sequenced Eurasian samples. In striking contrast to expectations, we find that all African individuals sampled are native to this continent, including those from sub-Saharan Africa. Moreover, we show that Africa harbors the greatest variation and represents the deepest history in the *A. thaliana* lineage. Our results also reveal evidence that selfing, a major defining characteristic of the species, evolved in a single geographic region, best represented today within Africa. Demographic inference supports a model in which the ancestral *A. thaliana* population began to split by 120–90 kya, during the last interglacial and Abbassia pluvial, and Eurasian populations subsequently separated from one another at around 40 kya. This bears striking similarities to the patterns observed for diverse species, including humans, implying a key role for climatic events during interglacial and pluvial periods in shaping the histories and current distributions of a wide range of species.

evolution | population history | self-compatibility | climate | migration

The plant *Arabidopsis thaliana* is the principal plant model species, and as such has been useful not only to examine basic biological mechanisms but also to elucidate evolutionary processes. The exceptional resources available in this species, including seed stocks collected from throughout Eurasia for over 75 y, have been a valuable tool for learning about the natural history of *A. thaliana* on this continent (1, 2). Previous studies have shown that current variation in Eurasia is mainly a result of expansions and mixing from refugia in Iberia, Central Asia, and Italy/Balkans after the end of the last glacial period ∼10 kya (3–8). The main finding of the recent analysis of 1,135 sequenced genomes was that a few Eurasian samples represent divergent relict lineages, whereas the vast majority derived from the recent expansion of a single clade (4). Given the large number of studies that examine the natural history of *A. thaliana*, one would expect that this history would by now be described rather completely and there would be no major surprises left to uncover. However, there are still many open questions about the ancient history of the species.

Several features differentiate *A. thaliana* from its closest relatives. Although most members of the *Arabidopsis* genus are obligate outcrossing perennials with large flowers and genome sizes of over 230 Mb and 8 chromosomes, *A. thaliana* is a predominantly selfing annual with reduced floral morphology and a reduced genome size of ∼150 Mb and 5 chromosomes. The transition to predominant selfing in *A. thaliana* was likely the catalyst for these derived morphological and genomic features (9–13). These changes, in particular the rearranged and shrunken genome, created a strong reproductive barrier between *A. thaliana* and its closest relatives (14).

Although the genetic basis of self-compatibility in *A. thaliana* is known, the specific events that occurred during the transition to predominant selfing are still unclear. In obligate out-crossing *Arabidopsis* species, many highly divergent S-locus haplogroups (S-haplogroups) are maintained by balancing selection, providing a mechanism for inbreeding avoidance. In *A. thaliana*, three S-haplogroups are found, and each contains mutations that obliterate function of the S-locus genes (15–17). Loss-of-function occurred independently in each S-haplogroup (18–21), but because these three S-haplogroups were never found together in the same geographic region, self-compatibility is inferred to have evolved separately in multiple locations (16, 21, 22). However, the hypothesis of geographically distinct origins is difficult to reconcile with the major genomic and phenotypic changes that render *A. thaliana* incompatible with its out-crossing congeners (9–13). Shifts from out-crossing to predominant selfing are common and have been considered the most prevalent evolutionary transitions in flowering plants (23). Reconstructing the evolutionary history of the transition to selfing in *A. thaliana* could provide general insights into this common evolutionary

Significance

The principal plant model species, *Arabidopsis thaliana*, is central to our understanding of how molecular variants lead to phenotypic change. In this genome-sequencing effort focused on accessions from Africa, we show that African populations represent the most ancient lineages and provide new clues about the origin of selfing and the species itself. Population history in Africa contrasts sharply with the pattern in Eurasia, where the vast majority of samples result from the recent expansion of a single clade. This previously unexplored reservoir of variation is remarkable given the large number of genomic studies conducted previously in this well-studied species and implies that assaying variation in Africa may often be necessary for understanding population history in diverse species.

Author contributions: A.D., A.F., and A.M.H. designed research; A.D., A.F., R.M.G., S.I.A., P.J.F., C.N., T.T., H.A.B., and A.M.H. performed research; A.D., A.F., F.X.P., C.A.-B., and A.M.H. contributed new reagents/analytic tools; A.D., A.F., and A.M.H. analyzed data; and A.D., A.F., and A.M.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper has been deposited in the European Nucleotide Archive/Sequence Read Archive database, study PRJEB19780 (accession nos. ERS1575066–ERS1575147). Analysis scripts are available at https://github.com/HancockLab/African_A.thaliana.

[1]A.D. and A.F. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: hancock@mpipz.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1616736114/-/DCSupplemental.

transition. However, because substantial time has passed since this transition (24–26) and no intermediate forms have been found between *A. thaliana* and its obligate out-crossing relatives, this reconstruction is challenging.

We sequenced the genomes of 78 African samples and analyze these in combination with 1,135 previously sequenced samples (4) (Fig. 1 and *SI Appendix*, Table S1). We find that African variation reveals the ancient history of the species and clarifies details concerning the transition to selfing. Congruence of *A. thaliana* population history with major climatic events and paleontological observations illustrates the relevance of population genetic studies for understanding climate-mediated demography more generally.

Results

To examine the relationship between African individuals and other worldwide samples, we used three complementary clustering approaches. Distance-based clustering by neighbor-joining reveals a clear split between Eurasian and African samples, indicating deep divergence between the continents (Fig. 2*A* and *SI Appendix*, Fig. S1). The majority of Eurasian samples form a nearly star-shaped phylogeny, consistent with recent expansion of these lineages. Conversely, longer more bifurcated branches separate African subclusters and previously identified Eurasian relicts from each other and from the nonrelict clade. In general, the Eurasian clades cluster consistently with the nine groups defined previously (4). Exceptions are the Central Europe clade, which separates into two clusters, and the Iberian relicts, which cluster with the Moroccan Rif-Zin population. Moroccan samples separate into four clades, reflecting their geographic distribution and South Africa and Tanzania cluster together in a single clade. The results for South Africa and Tanzania are striking because *A. thaliana* populations outside of Eurasia and North Africa were previously thought to be recently introduced by humans (27).

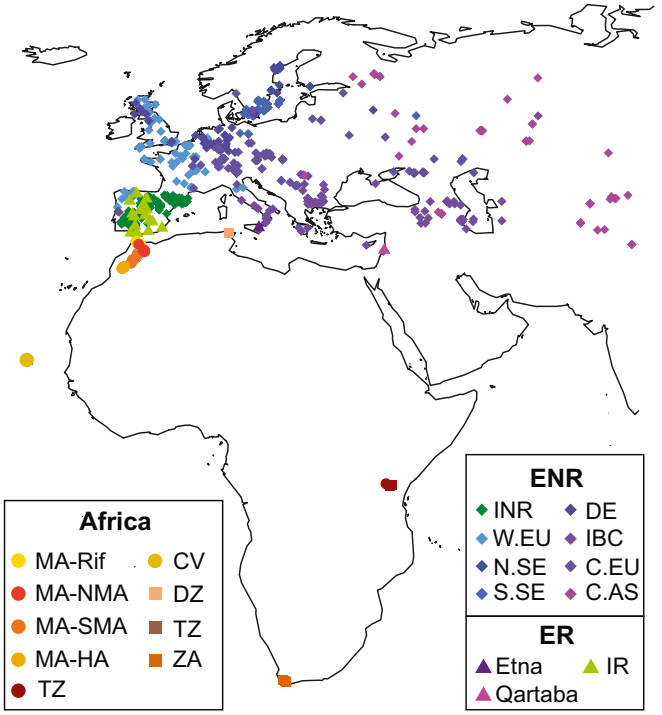


Fig. 1. Sample map of accessions included in this study. Herbarium samples are shown as squares. Abbreviations are as follows: Algeria (DZ), Cape Verde (CV), Central Asia (C.AS), Central Europe (C.EU), Eurasian nonrelicts (ENR), Eurasian relicts (ER), Germany (DE), Italy, Balkans, and Caucasus (IBC), Iberian nonrelicts (INR), Iberian relicts (IR), Morocco (MA), North Sweden (N.SE), South Africa (ZA), South Sweden (S.SE), Tanzania (TZ), Western Europe (W.EU).

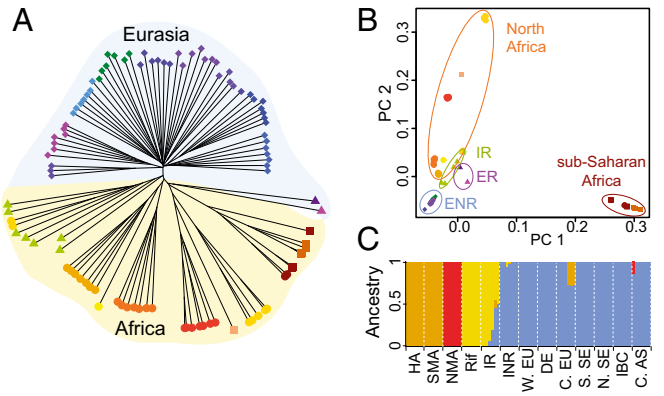


Fig. 2. Global population structure. (*A*) Unrooted neighbor-joining tree, (*B*) PCA, (*C*) ADMIXTURE results for $K = 4$.

Similarly, principal component analysis (PCA) separates African populations from each other and from Eurasian populations (Fig. 1*B*). The first PC distinguishes sub-Saharan Africa, and the second separates the four Moroccan clusters from Eurasians. Subsequent PCs mainly discriminate populations within Africa, whereas Eurasian populations remain tightly clustered (*SI Appendix*, Fig. S3). Results from ADMIXTURE (28) reinforce this finding. Moroccan populations separate into three clusters and are distinguished from a single cluster of Eurasian samples (Fig. 1*C*, and *SI Appendix*, Figs. S4 and S5, and Table S2). PCA and ADMIXTURE results also suggest a Moroccan origin of the relicts in Iberia, which are spread between Rif-Zin and Iberian nonrelicts in PCA, and sizable portions of the Iberian relict genomes match the Moroccan clusters in ADMIXTURE. This finding is consistent with previous work (29) and with the accepted phylogeographical history of Mediterranean and North African flora characterized by a complex history of expansions and contractions driven by important climatic changes experienced in this vast region, particularly since the Pliocene (30, 31).

Furthermore, from pairwise differences, we recover the previously reported difference between Eurasian relicts and nonrelicts (4) and find that all African accessions are at least as divergent as samples previously classified as relicts (Fig. 3*A* and *SI Appendix*, Fig. S6). Therefore, in contrast to Eurasia, where most samples represent a single recently spread clade, all African individuals represent relict samples. The distribution of pairwise differences within Africa (Fig. 3*B*) further demonstrates the high diversity in these samples.

If populations in Africa truly are more ancient than the Eurasian clusters, we should also expect higher numbers of private variants in Africa. Indeed, we find that the Moroccan clusters and Iberian relicts, which appear likely derived from Morocco, harbor the highest numbers of private SNPs (Fig. 3*C* and *SI Appendix*, Fig. S8 and Table S3). This signal intensifies when we exclude recently arisen variants, which we find constitute the majority of the private variation in Eurasia. First, we excluded singletons, the class of SNPs most influenced by recent population growth, and found a 7.0- to 23.2-fold enrichment in Morocco compared with the top nonrelict cluster (Fig. 3*C*). Next, we considered the spatial distribution of private variants across the genome. Because novel private variants are unlikely to be tightly linked, clustering of several, contiguous private SNPs indicates old haplotypes. At the haplotype level, we again found extremely high enrichment (3.9- to 20.9-fold) for Moroccan clusters and Iberian relicts (4.5-fold) relative to the top Eurasian cluster (Fig. 3*C*). Notably, the Eastern

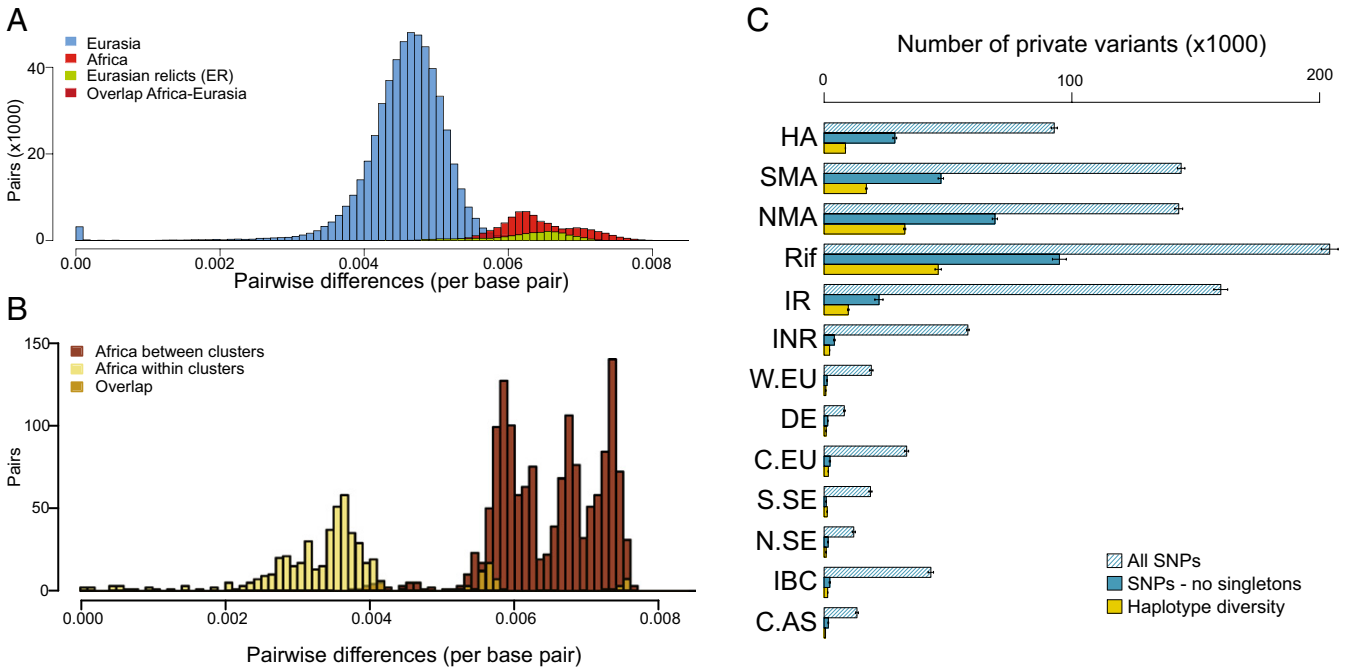


Fig. 3. Patterns of diversity across geographic regions. Distributions of genome-wide pairwise differences per base pair in: (*A*) worldwide comparison and (*B*) within and between African populations, where overlap between distributions is shown as described in legends. (*C*) Numbers of private SNPs and haplotypes found in each cluster. Error bars denote 95% confidence intervals.

Mediterranean and Caucasus regions, which were previously favored as points of origin and major centers of diversity of the species (3), do not exhibit a striking pattern for any of the metrics examined (see IBC in Fig. 3*C*). These findings evoke a model in which polymorphism in Africa is because of ancient variation and Eurasian polymorphism is mainly because of recent expansion.

To better understand the relevance of variation in Africa for the early history of the *A. thaliana* lineage, we examined variation at the locus that confers self-compatibility. S-locus variation in Africa differs in several ways from what is found in Eurasia (Fig. 4 and *SI Appendix*, Table S4). We found all three S-haplogroups together in a single geographic region, with S-haplogroup B private to Africa. In addition, the A-C recombinant is also present at low frequency in Morocco, in contrast to what is found in Eurasia (32). Finally, we discovered deletion haplotypes in haplogroups A and C in Morocco (*SI Appendix*, Fig. S9). The finding that all S-locus haplogroups are present together implies that selfing evolved in a single geographic region.

Taken together, the patterns in population structure, and levels of variation across the genome and at the S-locus, specifically, imply a deep history in Africa. To clarify the details of past demographic events, we inferred historical effective population sizes (N_e) and split times among populations based on cross-coalescent rates (CCR) using a multiple sequentially Markovian coalescent approach (MSMC) (33).

In ancient times, we find the highest N_e is in Africa, peaking at around 500–400 kya (Figs. 5 and *SI Appendix*, Fig. S10). All Eurasian populations (including Iberian relicts) show the same trajectories as the Africans, but with lower amplitudes. Given that these curves are in phase with one another and we do not see evidence for a population split until 120 kya (Fig. 6*A*), we interpret this as population structure in the ancestral population combined with bottlenecks as more derived populations migrated away from this ancestral population. This finding is consistent with our finding that variation in Eurasia is often a subset of variation present within Africa and is similar to the situation in humans (34). Notably, the IBC cluster, which includes previously hypothesized *A. thaliana* origins and refugia (3, 5, 7), exhibits a much lower ancient population size than Africa (*SI Appendix*, Fig. S10).

At 120–90 kya, there are bottlenecks in all populations (Fig. 5) and a split among the major clades (Moroccan, Tanzanian, and Levant) (Fig. 6 and *SI Appendix*, Fig. S11). This roughly corresponds to the Abbassia Pluvial, a period when migration corridors were open because of high precipitation and humidity in Africa (35) and also marks the last interglacial at Marine Isotope Stage 5e (130–116 kya), when temperatures were 1–2° warmer than present-day conditions, providing favorable conditions in Eurasia. As this interglacial period came to a close, there was a worldwide shift toward cooler, drier conditions as the most recent and severe Pleistocene glaciation phase began (36). Beginning at this time, CCR implies a progressive decline in population connectivity, consistent with decreasing temperature and increasing aridity (Fig. 6*A*). We checked for consistency using a complementary method that relies on the joint site frequency spectrum between populations ($\delta a \delta i$) (37) and found a slightly older estimate for the split and overlapping confidence intervals (141–116 kya) (*SI Appendix*, Table S5). We propose that the most likely scenario is that *A. thaliana* was colonizing broadly within Africa as well as in the Levant during the last interglacial (130–116 kya), but that connections between populations began to break down as populations spread and as climate became cooler and drier at the end of this period.

MSMC based on eight haplotypes detects several changes in N_e in more recent times (Fig. 5*B*). Since ~120 kya, the population size changes in Europe and Asia are often out of phase with those in Africa, consistent with geographical separation and exposure to different climatic regimes. Maxima in African populations occur at around 60–40 kya and 11–5 kya, corresponding to orbital-scale climate shifts from arid to moist conditions (i.e., pluvial periods) that occurred at 11–5 kya and 59–47 kya in North Africa (38, 39). We find relatively recent split times between sub-Saharan African clades (South Africa and Tanzania) and between Western Europe and Central Asia, at around 40 kya (Fig. 6).

There are a few caveats to consider regarding the demographic inference. Because MSMC can spread instantaneous population

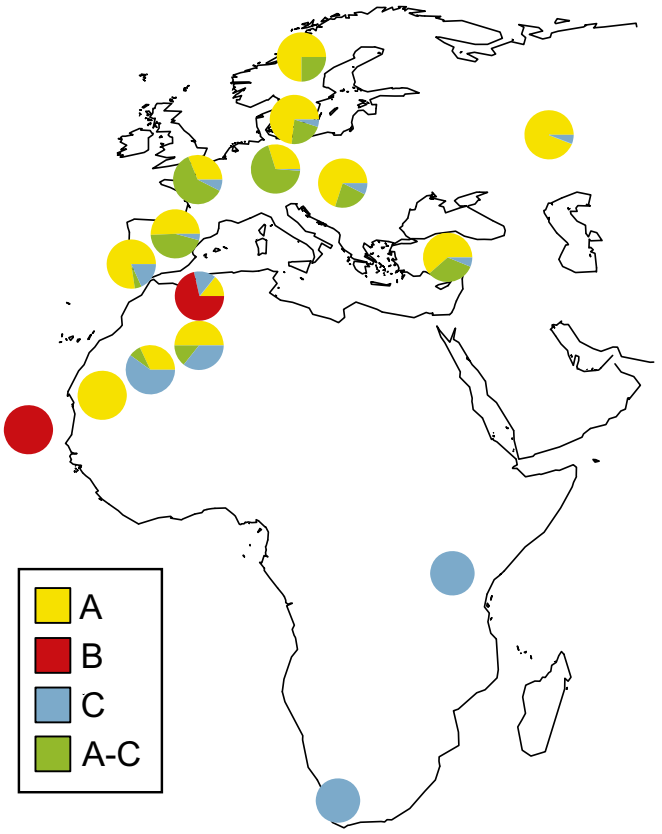


Fig. 4. Map of S-locus haplogroup diversity.

size changes over time, maxima and minima are informative but the slope should be interpreted with caution (33). In addition, the precise timing associated with inferred population size changes and splits is dependent on parameters that are difficult to measure and may vary over space and time, including mutation rate, degree of purifying selection, and the possible input from a seed bank. We used the best available data for mutation rate [based on mutation accumulation experiments (40)] and made the usual simplifying assumptions for other parameters (one generation per year), but the timing we infer would need to be revised if these assumptions turned out to be incorrect.

Discussion

Genomic studies thus far have amassed data for nearly 2,000 Eurasian *A. thaliana* accessions but were unable to provide insight into the early history of the species. Here, in a genome-scale sequencing effort focused on African accessions, we find clear evidence for a deep history of African *A. thaliana* populations, which harbor variation that was either lost or never present in Eurasia. Several specific results were unexpected based on current knowledge in this well-studied species. First, we discovered surprising and clear evidence that *A. thaliana* is native not only to North Africa but also to Afro-alpine regions of sub-Saharan Africa. Second, our results revealed that the deepest splits species-wide separate the African lineages from one another and that in ancient times, the effective population size was largest in Africa. Finally, we learned that variation at the S-locus is highest in Africa and that all three S-haplogroups are present there.

Based on our results, we can outline a model for the early history and transition to selfing in *A. thaliana* (detailed in *SI Appendix*, Fig. S12). In the first step, we infer that the population ancestral to *A. thaliana* became geographically separated from its parental out-crossing population. Our results suggest that this separation involved migration of the ancestral subpopulation into Africa by 1.2–0.8 Mya. This timing corresponds to the Middle Pleistocene Transition, a shift to drier more variable climate and more open habitats in Africa (i.e., grasslands versus woodlands), as evidenced by soil carbon analysis showing an increase in the ratio of C4 to C3 plants (41, 42).

Although the estimated divergence times between *A. thaliana* and *Arabidopsis lyrata* center around 5–7 Mya (9, 43), the origin of *A. thaliana* itself appears to be much younger. Our model predicts that there was an initial bottleneck as the subpopulation that led to *A. thaliana* split from a *A. lyrata*-like ancestral population [similar to that observed in *Mimulus nasutus* (44) and *Capsella rubella* (45–47)], followed by an expansion in N_e as the selfing population began to spread. In this case, we could interpret the MSMC results to suggest that the transition to selfing occurred between 1 Mya and 500 kya, before the most ancient maximum in N_e. This finding is in line with an estimate based on the depth of the *A. thaliana* genealogy (0.84% maximum divergence among individuals sampled here) under a simple model ($T \sim D/2\,\mu \sim 598$ kya). Our estimated timing is also consistent

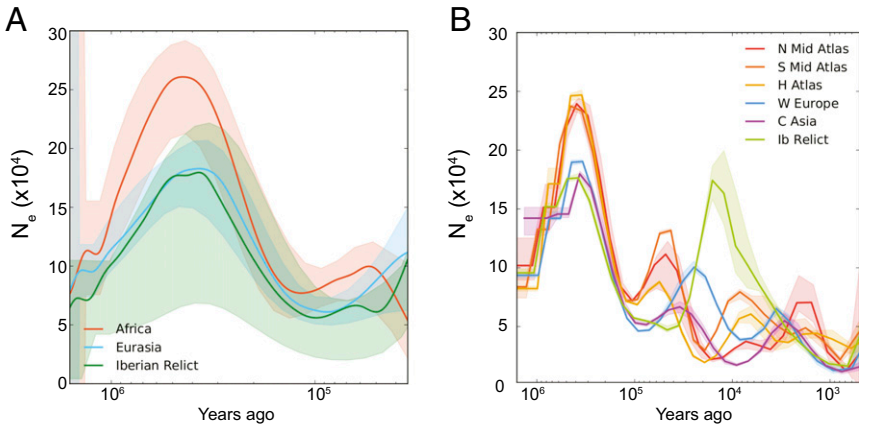


Fig. 5. Historical effective population size of *A. thaliana* inferred using MSMC. Although two-haplotype analysis provides more resolution in the distant past, eight-haplotype analysis provides better resolution in the recent past. (*A*) Inference using pairs of haplotypes, with lines representing medians and shading representing ±1 SD calculated across pairs. This analysis is expected to produce unbiased estimates between 40 kya and 1.6 Mya (*SI Appendix*). (*B*) Inference based on sets of eight haplotypes with lines representing medians. This analysis is expected to produce unbiased estimates as recently as 1.6 kya.

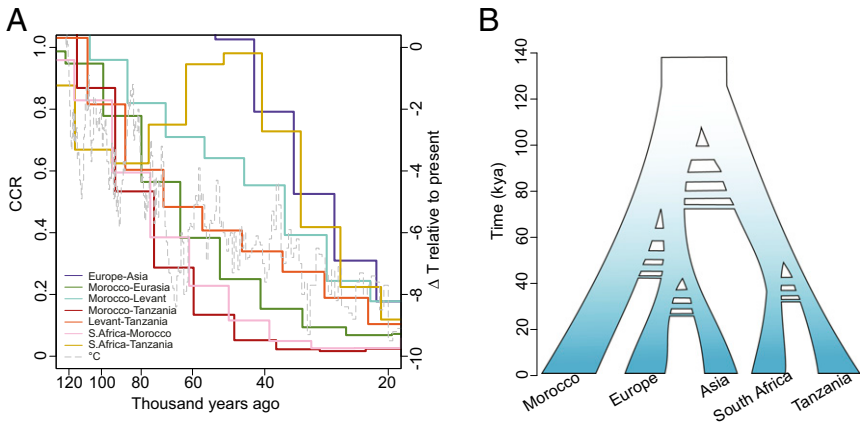


Fig. 6. Inferred timing of population splits. (*A*) Relative CCR between populations. Decreasing values from 1.0 indicate population separation. The dashed line represents historical temperature (63). (*B*) A schematic model for the demographic history of *A. thaliana* based on CCR results, with hashes to represent uncertainty regarding possible timing of gene flow events.

with previous estimates for the loss of self-incompatibility and origin of selfing (24–26).

Once selfing was established, traits associated with the "selfing syndrome" would have been favored, including reduced pollen number and petal size (48). Such phenotypic shifts are common in predominantly selfing species and have occurred in *A. thaliana* compared with its closest relatives (26). At the genomic level, *A. thaliana* exhibits major chromosomal rearrangements and a reduction in genome size and number of chromosomes (49). This genomic reduction is also likely a by-product of the shift to predominant selfing in *A. thaliana* (9–11), consistent with an observed link between reduced genome size and selfing in other plant species (11–13). These changes introduce a strong reproductive barrier as found in hybrids of *A. thaliana* and *A. lyrata*, which are infertile because of the chromosomal rearrangements that occurred in *A. thaliana* (14).

Given that all three S-haplogroups co-occur in Morocco, we hypothesize that the transition to predominant selfing occurred in a single region, best represented today in Morocco. This finding differs from previous assertions that these events likely happened separately in geographically distinct populations (15, 16). Moreover, it allows for the possibility that the transition to selfing was aided by a shared precursor mutation, a shared climate, and the bottleneck that occurred during the migration away from the ancestral population (22). Our proposed model parallels observations in partially selfing populations of *A. lyrata* (50). Here, self-compatibility is associated with two different S-haplogroups in Great Lakes populations and self-compatibility may have been favored because of the bottleneck that initially limited S-haplogroup diversity and thus mate availability.

After the origin and initial population size increase of *A. thaliana*, we infer several demographic changes that are congruent with known climatic shifts. At 120–90 kya, we find evidence from MSMC and *δaδi* for splitting among the major clades: Morocco, Levant and sub-Saharan Africa. This split corresponds to the Abbassia pluvial, which produced migration corridors within Africa (120–90 kya) (35, 39) as well as Marine Isotope Stage 5e (130–116 kya), the last interglacial period, when worldwide temperatures were 1–2° warmer than they are currently (51, 52). This is consistent with a model in which *A. thaliana* spread widely throughout Africa and into Eurasia when conditions were favorable (~120 kya), with isolation as gene flow was reduced (*SI Appendix*, Fig. S12). More recent major demographic events include the split between European and Asian populations at around 40 kya and the increase in N_e within Africa during the most recent pluvial.

The patterns we observe and their concordance with climatic events suggest that the transition to selfing and speciation occurred within Africa, with subsequent migration out of Africa into Eurasia. However, it is also possible that the initial transition to selfing occurred within Eurasia followed by migration into Africa and concomitant loss of variation in Eurasia. This alternative would require that the ancient variation in the *A. thaliana* lineage was either lost or has not been sampled in Eurasia and the bottleneck into Africa was mild enough to preserve high levels of genetic variation.

Overall, the patterns in *A. thaliana* bear striking similarities to those observed for human populations, particularly in the larger effective population size in Africa (34), the exodus from Africa approximately 120 kya (39, 53–55), and the splitting of major human populations in Europe and Asia (approximately 45–35 kya) (53, 54). Analogous to what we propose here, demographic events in human populations have been attributed to major climate transitions (35, 39, 56).

Moreover, the timing and types of demographic events we infer during the history of *A. thaliana* are consistent with previous observations in a broad range of other plant species. Specifically, the shift to predominance of C4 plants across Africa at 1.2–0.8 Mya and the intensification of glacial cycles worldwide (57) correspond with our estimated timing of the evolution of selfing in *A. thaliana* and a clustering of speciation events more generally (58). The geographic expansion approximately 120 kya corresponds to an African pluvial and worldwide interglacial, which resulted in expansion of forests across Africa (59) and Eurasia (51, 52). Finally, we see evidence of an increase in effective population size overlapping with the most recent and well-described African pluvial at 11–5 kya, when the Sahara was heavily vegetated and filled with lakes (38, 60, 61). The concordance between inferred population size changes, climate, and reports for other species implies that the patterns we observe in *A. thaliana* may be representative of climate-mediated population dynamics across diverse taxa.

Materials and Methods

For full materials and methods, please see *SI Appendix, Supplementary Text*.

We sequenced the genomes of 79 *A. thaliana* individuals, including 70 fresh samples and 9 herbarium samples (*SI Appendix*, Table S1). For fresh leaf samples, sequencing libraries were prepared using Illumina TruSeq DNA sample prep kits (Illumina) and sequenced on Illumina Hi-Seq instruments. DNA from herbarium specimens was extracted, authenticated, and treated with uracil glycosylase to remove damaged nucleotides in a clean room facility at the University of Tübingen. To align the sequences to the TAIR10 reference genome and to call variants, we used two different pipelines: the MPI-SHORE pipeline (62) and a more conservative pipeline designed to reduce false positives resulting from indels.

For population structure analyses, we subsampled the complete dataset to match sample sizes across clusters as some Eurasian geographic regions are heavily oversampled, which could cause biases in some analyses, and we pruned SNPs based on LD to select a representative set. For ADMIXTURE, the number of clusters (K) was determined based on the outcome of cross-validation analyses.

To infer patterns of effective population size and population separations over historical time, we used a MSMC v2 (33). Because *A. thaliana* accessions are inbred, we created pseudodiploids by combining chromosomes from pairs of individuals from the same populations and ran MSMC in the two- and eight-haplotype configurations (Fig. 5). We assumed a mutation rate of 7.1×10^{-9} based on results of mutation accumulation experiments (40) and a generation time of 1 y. We confirmed inferences using *δa δi* (37) on joint site frequency spectra from pairs of populations.

ACKNOWLEDGMENTS. We thank the South African National Biodiversity Institute, the University of Vienna, and Naturalis for access to herbarium samples; and V. Castric and M. Horton for helpful comments on the manuscript. Support for sequencing and plant growth were provided by the Vienna Biocenter Core Campus Science Support Facilities and the Max Planck Gesellschaft (MPG). Computational resources were from the Vienna Scientific Cluster, Center for Integrative Bioinformatics Vienna, and MPG were used. The project was supported by European Union CIG 304301, European Research Council Grant CVI_ADAPT 638810, and MPG (to A.M.H.); Austrian Science Funds FWF W1225; Grants BIO2016-75754-P (to C.A.-B.) and CGL2016-77720-P (to F.X.P.) from the Agencia Estatal de Investigación of Spain and the Fondo Europeo de Desarrollo Regional; Japan Society for the Promotion of Science KAKENHI 15K18583 (to T.T.); and the Presidential Innovation Fund of the MPG (to H.A.B.).

1. Provart NJ, et al. (2016) 50 years of *Arabidopsis* research: Highlights and future directions. *New Phytol* 209:921–944.
2. Somerville C, Koornneef M (2002) A fortunate choice: The history of *Arabidopsis* as a model plant. *Nat Rev Genet* 3:883–889.
3. 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* 17:902–915.
4. Consortium G; 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at; 1001 Genomes Consortium (2016) 1135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166:481–491.
5. François O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genet* 4:e1000075.
6. Picó FX, Méndez-Vigo B, Martínez-Zapater JM, Alonso-Blanco C (2008) Natural genetic variation of *Arabidopsis thaliana* is geographically structured in the Iberian peninsula. *Genetics* 180:1009–1021.
7. Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: Biogeography and postglacial colonization of Europe. *Mol Ecol* 9:2109–2118.
8. Lee C-R, et al. (2017) On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nat Commun* 8:14458.
9. Hu TT, et al. (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 43:476–481.
10. Oyama RK, et al. (2008) The shrunken genome of *Arabidopsis thaliana*. *Plant Syst Evol* 273:257–271.
11. Wright SI, Ness RW, Foxe JP, Barrett SCH (2008) Genomic consequences of outcrossing and selfing in plants. *Int J Plant Sci* 169:105–118.
12. Albach DC, Greilhuber J (2004) Genome size variation and evolution in Veronica. *Ann Bot (Lond)* 94:897–911.
13. Trivers R, Burt A, Palestis BG (2004) B chromosomes and genome size in flowering plants. *Genome* 47:1–8.
14. Nasrallah ME, Yogeeswaran K, Snyder S, Nasrallah JB (2000) *Arabidopsis* species hybrids in the study of species differences and evolution of amphiploidy in plants. *Plant Physiol* 124:1605–1614.
15. Sherman-Broyles S, et al. (2007) S locus genes and the evolution of self-fertility in *Arabidopsis thaliana*. *Plant Cell* 19:94–106.
16. Shimizu KK, Shimizu-Inatsugi R, Tsuchimatsu T, Purugganan MD (2008) Independent origins of self-compatibility in *Arabidopsis thaliana*. *Mol Ecol* 17:704–714.
17. Tsuchimatsu T, et al. (2010) Evolution of self-compatibility in *Arabidopsis* by a mutation in the male specificity gene. *Nature* 464:1342–1346.
18. Dwyer KG, et al. (2013) Molecular characterization and evolution of self-incompatibility genes in *Arabidopsis thaliana*: The case of the Sc haplotype. *Genetics* 193:985–994.
19. Liu P, Sherman-Broyles S, Nasrallah ME, Nasrallah JB (2007) A cryptic modifier causing transient self-incompatibility in *Arabidopsis thaliana*. *Curr Biol* 17:734–740.
20. Nasrallah JB (2002) Recognition and rejection of self in plant reproduction. *Science* 296:305–308.
21. Boggs NA, Nasrallah JB, Nasrallah ME (2009) Independent S-locus mutations caused self-fertility in *Arabidopsis thaliana*. *PLoS Genet* 5:e1000426.
22. Vekemans X, Poux C, Goubet PM, Castric V (2014) The evolution of selfing from outcrossing ancestors in Brassicaceae: What have we learned from variation at the S-locus? *J Evol Biol* 27:1372–1385.
23. Stebbins GL (1974) *Flowering Plants: Evolution Above the Species Level* (Harvard Univ Press, Cambridge, MA).
24. 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* 23:1741–1750.
25. Tang C, et al. (2007) The evolution of selfing in *Arabidopsis thaliana*. *Science* 317:1070–1072.
26. Shimizu KK, Tsuchimatsu T (2015) Evolution of selfing: Recurrent patterns in molecular adaptation. *Annu Rev Ecol Evol Syst* 46:593–622.
27. Hoffmann MH (2002) Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae). *J Biogeogr* 29:125–134.
28. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664.
29. 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.
30. Désamoré A, et al. (2011) Out of Africa: Northwestwards Pleistocene expansions of the heather *Erica arborea*. *J Biogeogr* 38:164–176.
31. Quézel P (1978) Analysis of the flora of Mediterranean and Saharan Africa. *Ann Mo Bot Gard* 65:479–534.
32. Tsuchimatsu T, et al. (April 4, 2017) Patterns of polymorphism at the self-incompatibility locus in 1,083 *Arabidopsis thaliana* genomes. *Mol Biol Evol*, 10.1093/molbev/msx122.
33. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46:919–925.
34. Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
35. 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 USA* 105:16444–16447.
36. Quante M (2010) The changing climate: Past, present, future. *Relict Species: Phylogeography and Conservation Biology*, eds Habel JC, Assmann T (Springer, Berlin).
37. 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:e1000695.
38. Tierney JE, deMenocal PB (2013) Abrupt shifts in Horn of Africa hydroclimate since the Last Glacial Maximum. *Science* 342:843–846.
39. Timmermann A, Friedrich T (2016) Late Pleistocene climate drivers of early human migration. *Nature* 538:92–95.
40. Ossowski S, et al. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94.
41. Cerling TE, Hay RL (1986) An isotopic study of paleosol carbonates from Olduvai Gorge. *Quat Res* 25:63–78.
42. deMenocal PB (2004) African climate change and faunal evolution during the Pliocene-Pleistocene. *Earth Planet Sci Lett* 220:3–24.
43. Hohmann N, Wolf EM, Lysak MA, Koch MA (2015) A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* 27:2770–2784.
44. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genet* 10:e1004410.
45. Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G (2013) Genomic identification of founding haplotypes reveals the history of the selfing species *Capsella rubella*. *PLoS Genet* 9:e1003754.
46. Foxe JP, et al. (2009) Recent speciation associated with the evolution of selfing in *Capsella*. *Proc Natl Acad Sci USA* 106:5241–5245.
47. Guo YL, et al. (2009) Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proc Natl Acad Sci USA* 106:5246–5251.
48. Sicard A, Lenhard M (2011) The selfing syndrome: A model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Ann Bot (Lond)* 107:1433–1443.
49. Johnston JS, et al. (2005) Evolution of genome size in Brassicaceae. *Ann Bot (Lond)* 95:229–235.
50. Mable BK, et al. (2017) What causes mating system shifts in plants? *Arabidopsis lyrata* as a case study. *Heredity (Edinb)* 118:110.
51. Kaspar F, Kühl N, Cubasch U, Litt T (2005) A model-data comparison of European temperatures in the Eemian interglacial. *Geophys Res Lett* 32:L11703.
52. Kukla GJ, et al. (2002) Last interglacial climates. *Quat Res* 58:2–13.
53. Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. *Proc Natl Acad Sci USA* 109:17758–17764.
54. Mallick S, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201–206.
55. Pagani L, et al. (2016) Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* 538:238–242.
56. deMenocal PB, Stringer C (2016) Human migration: Climate and the peopling of the world. *Nature* 538:49–50.
57. Tzedakis PC, Crucifix M, Mitsui T, Wolff EW (2017) A simple rule to determine which insolation cycles lead to interglacials. *Nature* 542:427–432.
58. deMenocal PB (1995) Plio-Pleistocene African climate. *Science* 270:53–59.
59. Dupont L (2011) Orbital scale vegetation change in Africa. *Quat Sci Rev* 30:3589–3602.
60. Cohmap M; COHMAP MEMBERS (1988) Climatic changes of the last 18,000 years: Observations and model simulations. *Science* 241:1043–1052.
61. Kuper R, Kröpelin S (2006) Climate-controlled Holocene occupation in the Sahara: Motor of Africa's evolution. *Science* 313:803–807.
62. Ossowski S, et al. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024–2033.
63. Kawamura K, et al. (2007) Northern Hemisphere forcing of climatic cycles in Antarctica over the past 360,000 years. *Nature* 448:912–916.

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

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

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

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

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

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

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

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

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

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

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

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

Effect of differences in coverage on diversity: To check whether genomic variation patterns are confounded by differences in coverage depth, we computed average coverage and diversity (average pairwise differences, θ_π) for each of the nine Eurasian and four Moroccan clusters. Moreover, we randomly subsampled reads in the raw data of all African samples, to the minimum average coverage across Eurasian clusters (15x, Fig. S7*A*), and repeated some of the analyses with this subsampled set. For this purpose we used the samtools 1.3.1 (12) <view –s> function, and we reprocessed all samples through the pipeline MPI-SHORE with the same parameters described above.
There was no significant correlation between average coverage and diversity, considering Eurasian clusters alone (Spearman rank correlation, r_s=-0.25, p=0.5165), Eurasian clusters together with Moroccan clusters (r_s=-0.27, p= 0.3737), and Eurasian clusters together with subsampled Moroccan clusters (r_s=-0.16, p=0.5916).
To check the effect of coverage depth at the level of single pairwise comparisons, we computed pairwise differences per base pair (diff./bp) for every pair of African samples (Fig. 2*B*) and for the same pairs after subsampling reads (Fig. S7*C*).

Artificially reducing coverage significantly lowered pairwise differences per base pair (paired $t(1829)= 127.0$, $p<0.01$). Nonetheless, this effect is of negligible magnitude (mean difference = 0.00026 diff./bp, and see Fig. S7*B*) compared to the actual differences across clusters (Fig. S7*A*, and compare Fig. 2*B* with Fig. S7*C*), so that divergence among African groups after reducing coverage (range: 0.361-0.737 % diff./bp, Fig. S7*C*) is of the same magnitude as before subsampling (range: 0.379-0.767 diff./bp %, Fig. 2*B*). Even after artificially reducing coverage, Moroccan clusters are as diverged from each other as Eurasian relicts from non-relicts (range: 0.356-0.748 % diff./bp).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SUPPLEMENTARY FIGURES

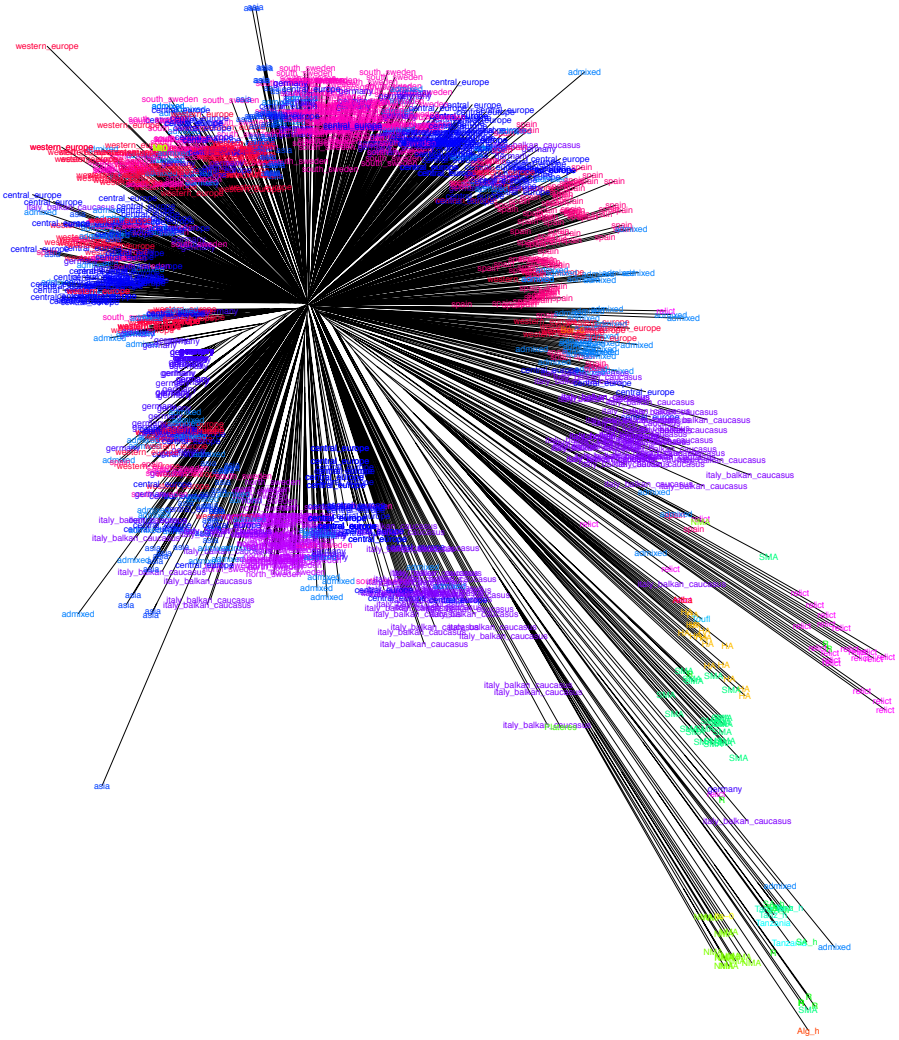


Figure S1. Neighbor-Joining tree with all samples, colored by cluster.

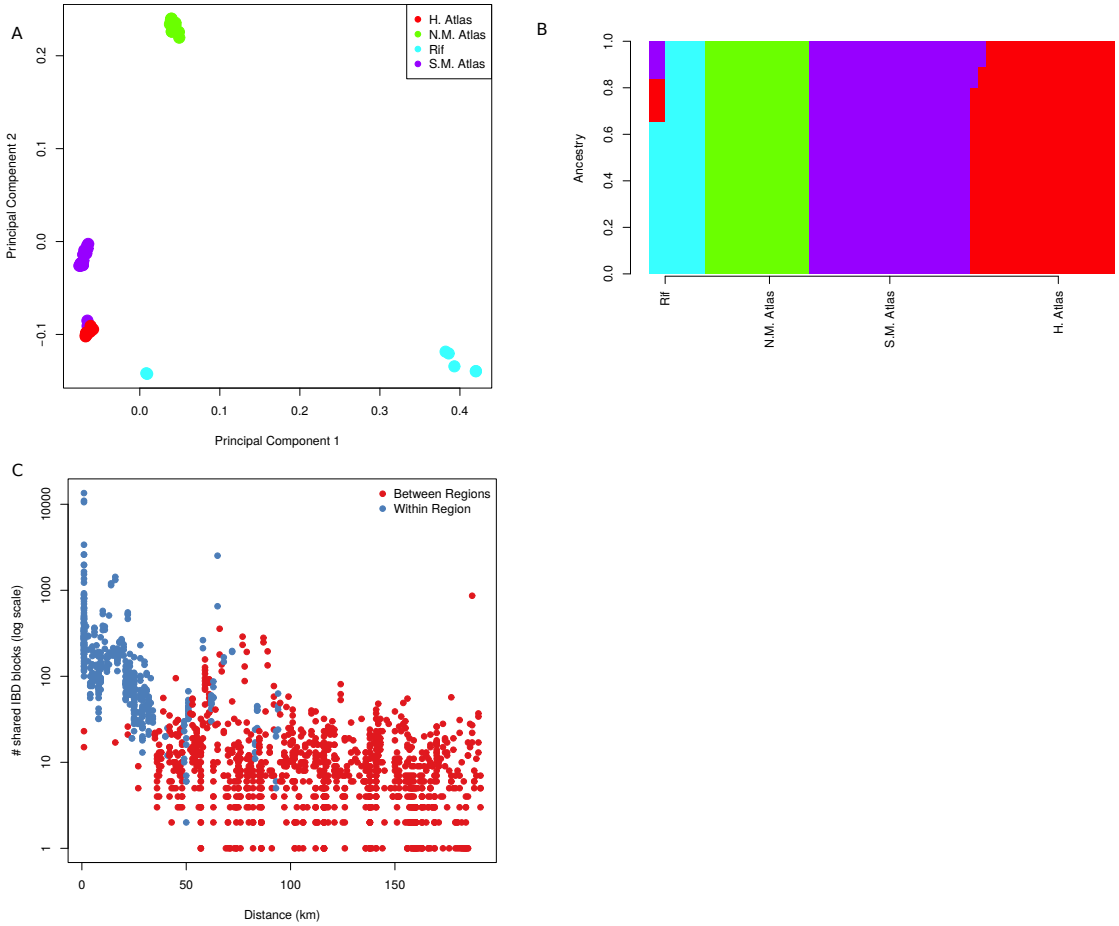


Figure S2. Population structure in Morocco.
A) Principal components analysis shows strong population structure in Morocco. B) ADMIXTURE with the lowest cross-validation error (K=4) confirms the strong population structure. C) Geographic decay of shared haplotypes calculated using RefinedIBD in Beagle.

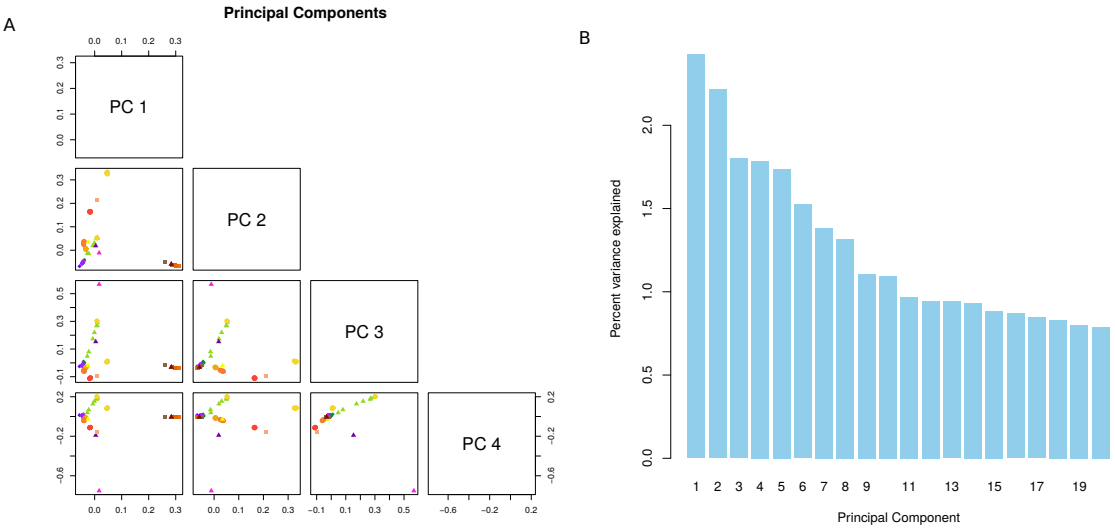


Figure S3. Worldwide PCA.
A) Worldwide PCA for the first 4 PCs. B) Proportion of variance explained by each PC up to PC 20.

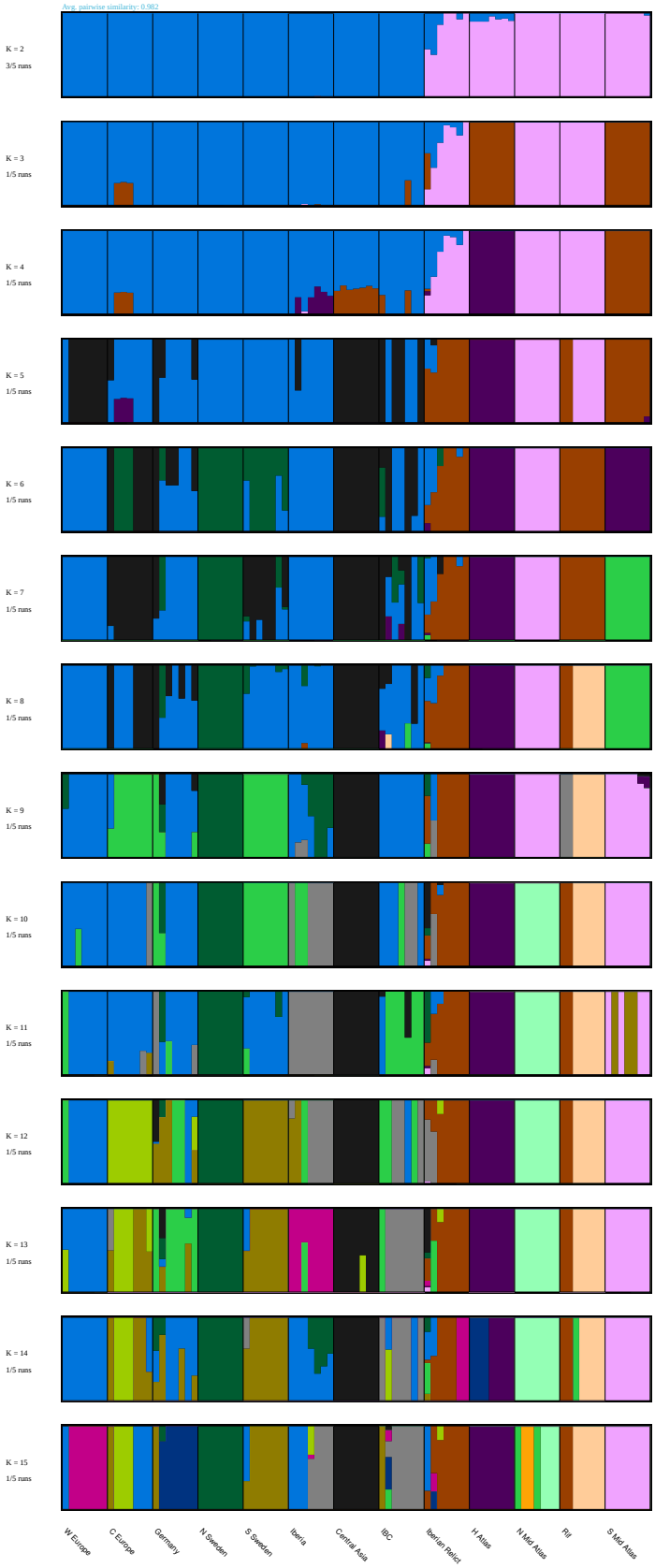


Figure S4. Worldwide ADMIXTURE for K=2 to K=15.
The run with the lowest cross-validation error (out of 20 replicates) is plotted.

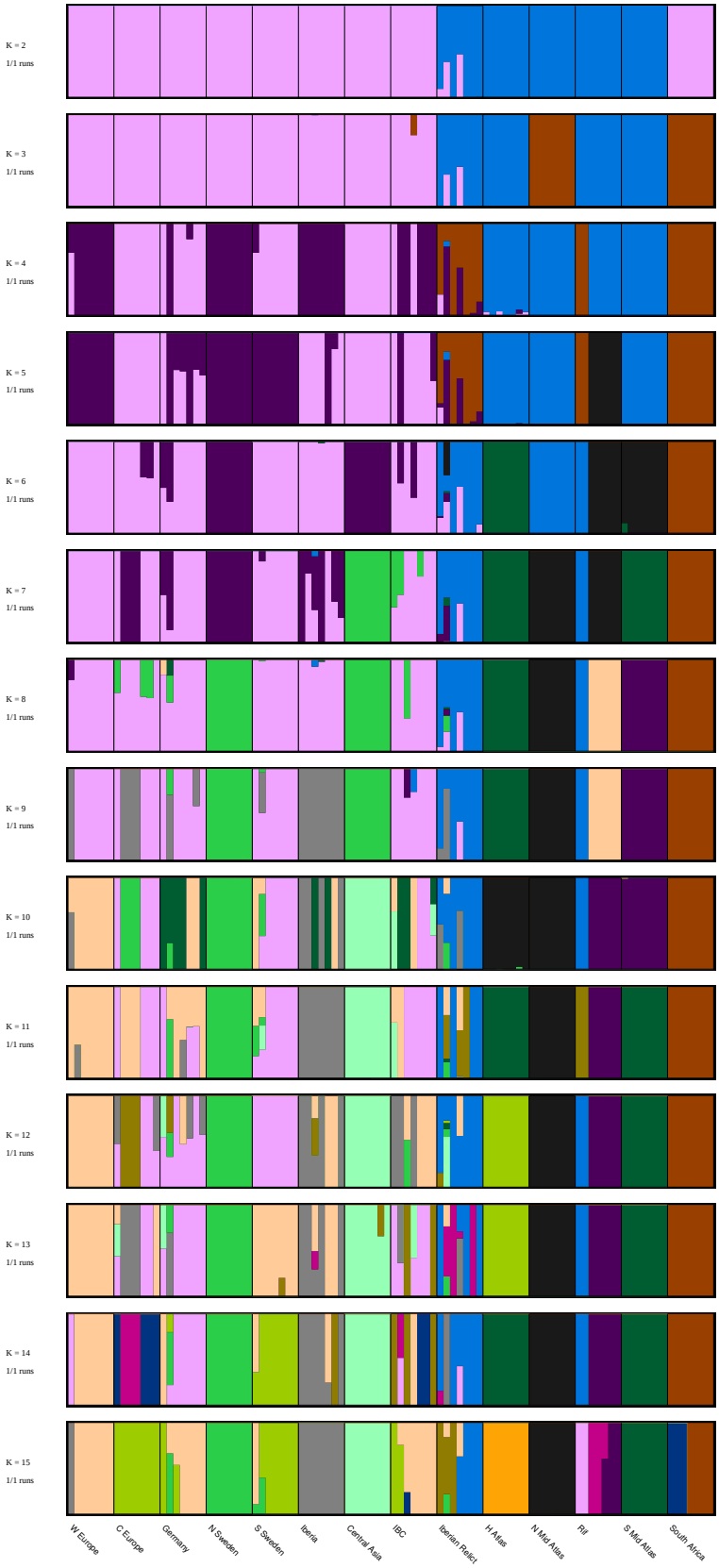


Figure S5. Worldwide ADMIXTURE including sub-Saharan samples for K=2 to K=15.

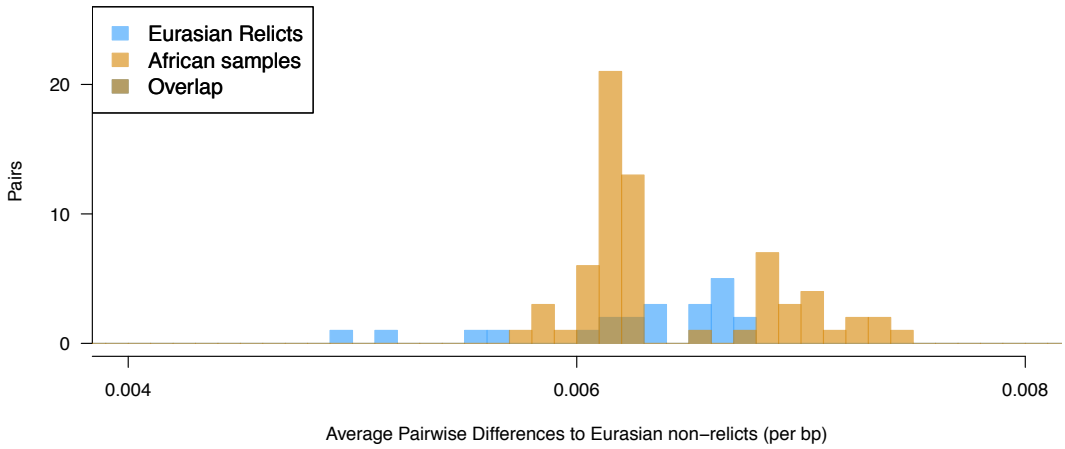
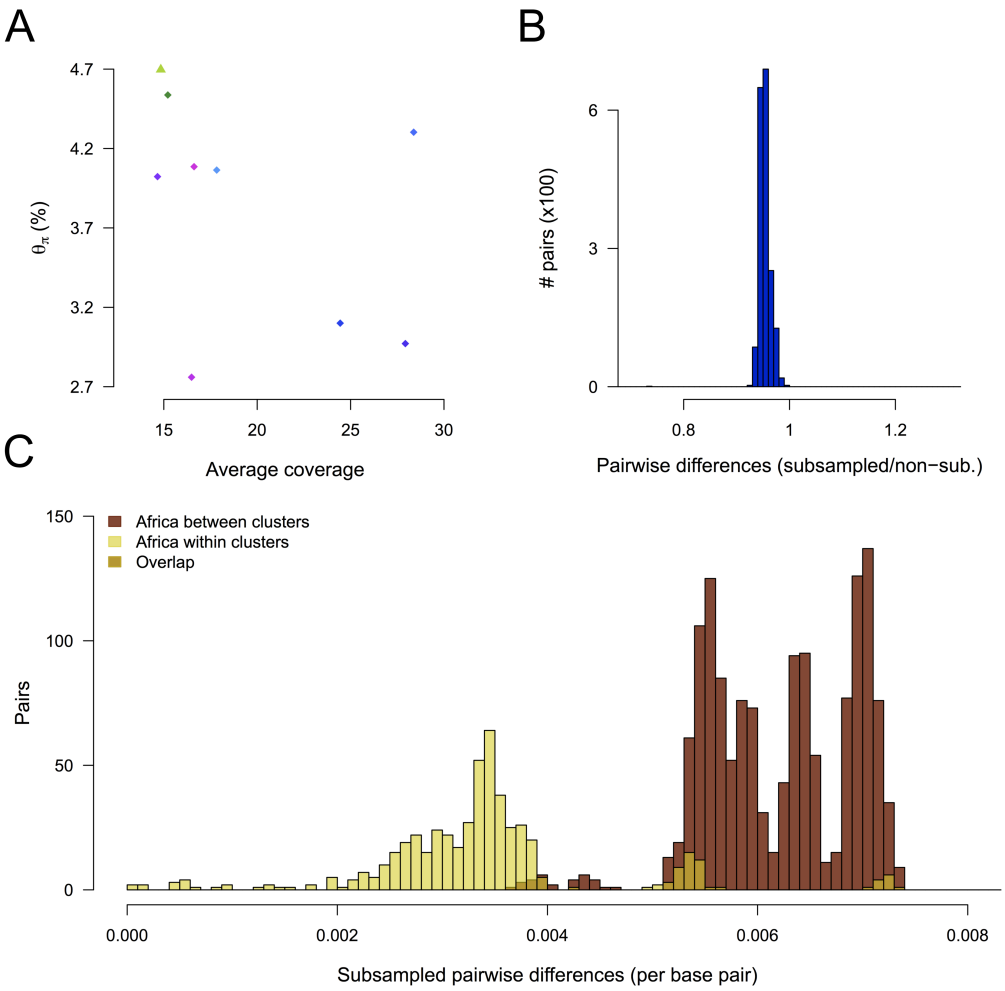
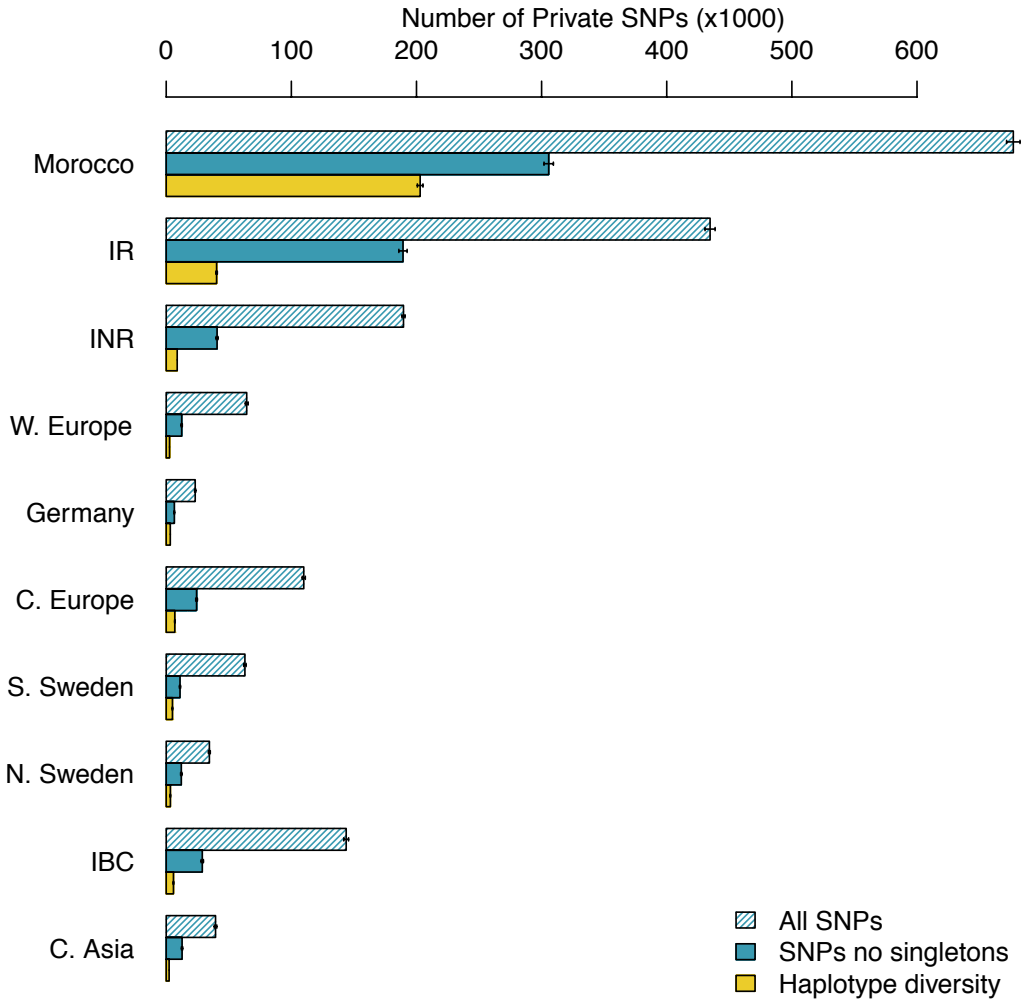


Figure S6. Divergence from Eurasian non-relicts.
Average pairwise differences was used as a measure of divergence of Eurasian relicts (blue) and Africans (tan) relative to Eurasian non-relicts. Every African accession meets the criterion for being defined as relict, and a subset of them are more diverged than any Eurasian relict.



406
407
408 **Figure S7. Effect of average coverage on diversity.**
409 a) Diversity within clusters (average pairwise differences, θ_π) as a function of average
410 coverage for all Eurasian clusters. There is no significant correlation between these
411 quantities. b) Ratio between pairwise differences among African samples after
412 artificially reducing coverage to 15x, the minimum average coverage among Eurasian
413 clusters (subsampled), and among full-coverage samples (non-sub.). c) Distribution of
414 pairwise differences per base pair for all African samples after reducing coverage to
415 15x (cfr. Fig. 2*B* for full-depth comparisons).
416
417



418
419
420 **Figure S8. Private variation with Morocco considered as a single clade.**
421 Numbers of private SNPs with singletons (hashed blue) and without singletons (dark
422 blue), and number of SNPs without singletons on private haplotypes (yellow) within
423 each cluster. Error bars show 95% confidence intervals from calculations across 500
424 resampled datasets. Morocco harbors the highest diversity in all measures of private
425 variation.
426
427

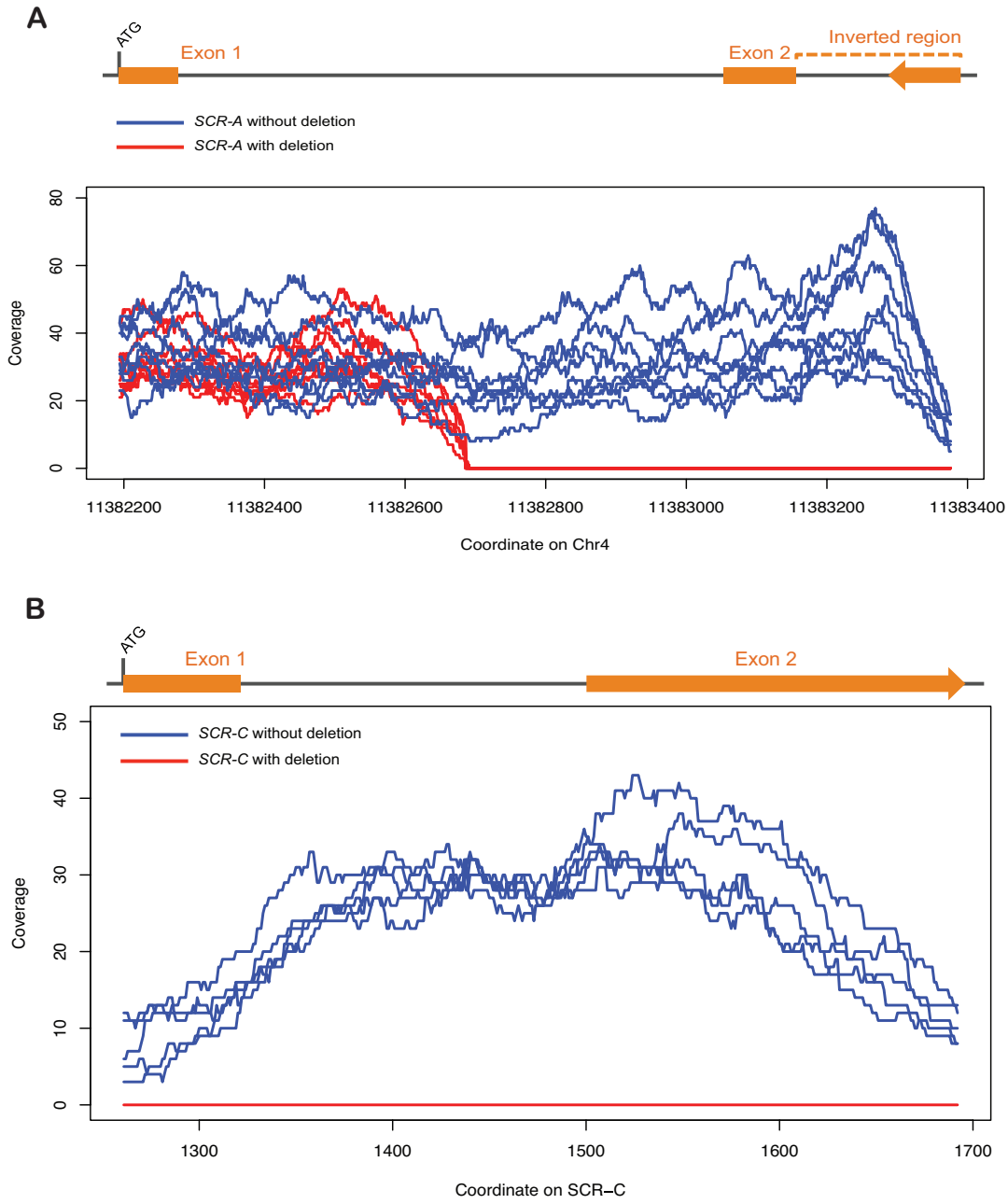
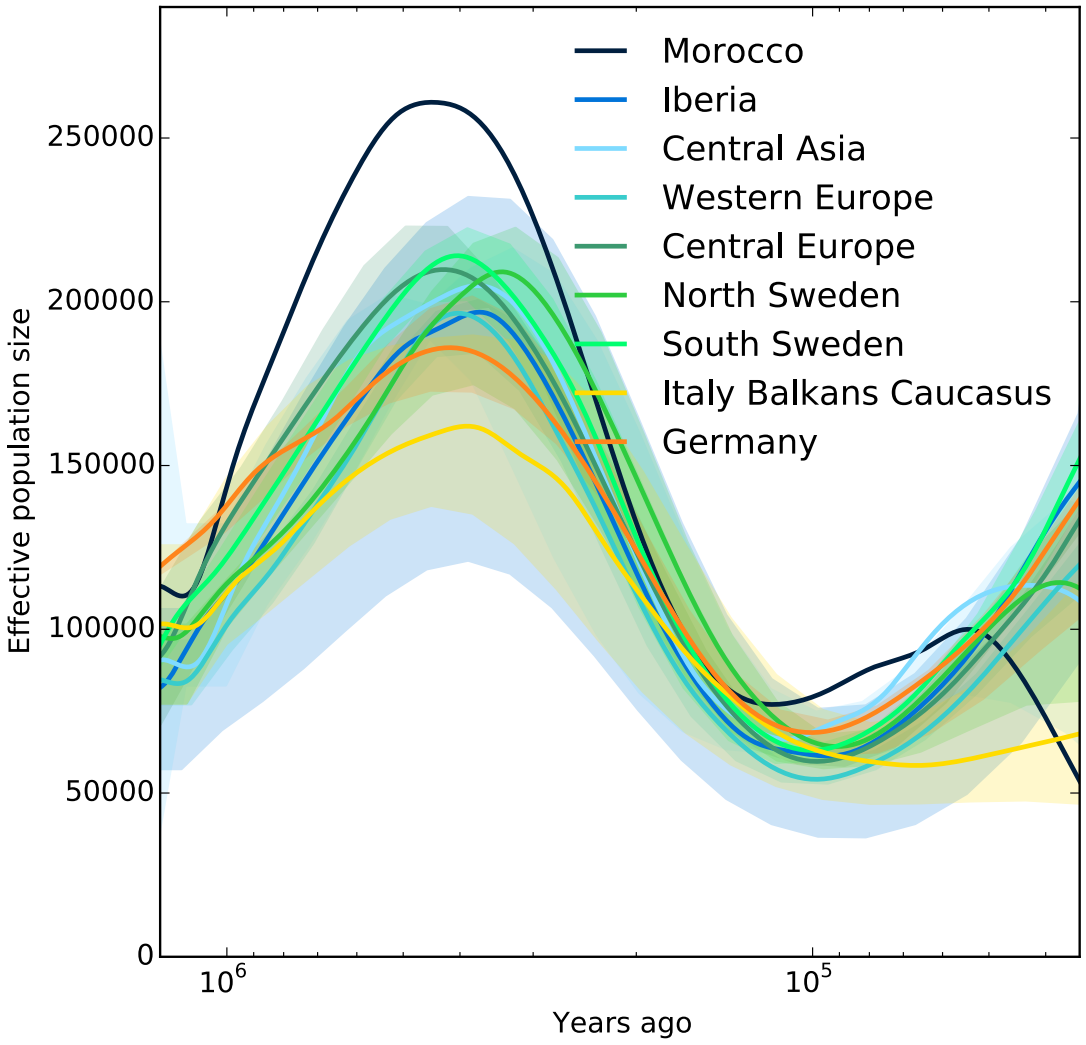


Figure S9. Novel deletion haplotypes for S-locus haplogroups A and C in Morocco.

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



436
437
438 **Figure S10. MSMC for Morocco and all Eurasian non-relict clusters.**
439 Effective population size medians are plotted with ± 1 standard deviation shaded.
440

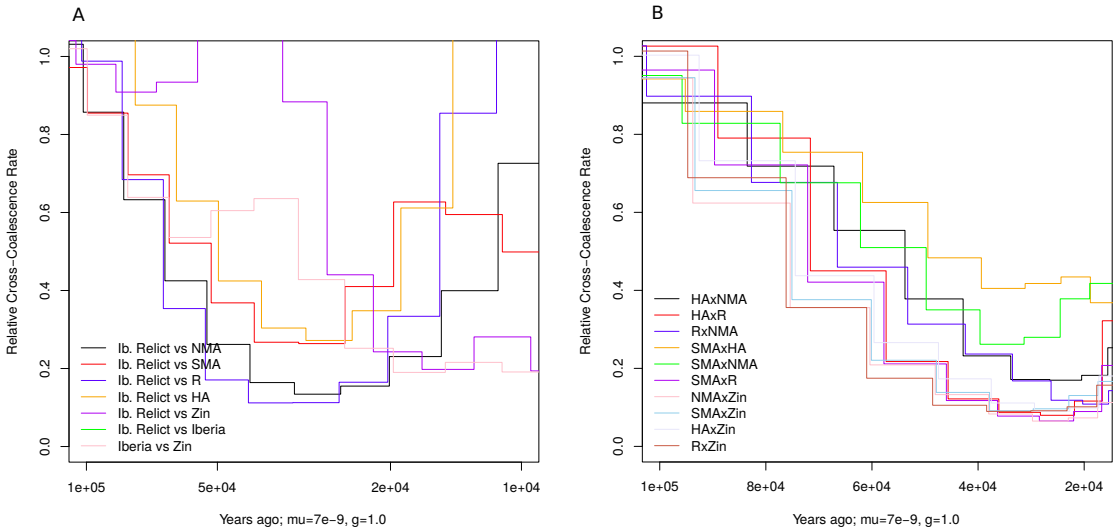


Figure S11. Relative cross coalescence for comparisons between regions within Morocco and Iberian Relicts.

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.

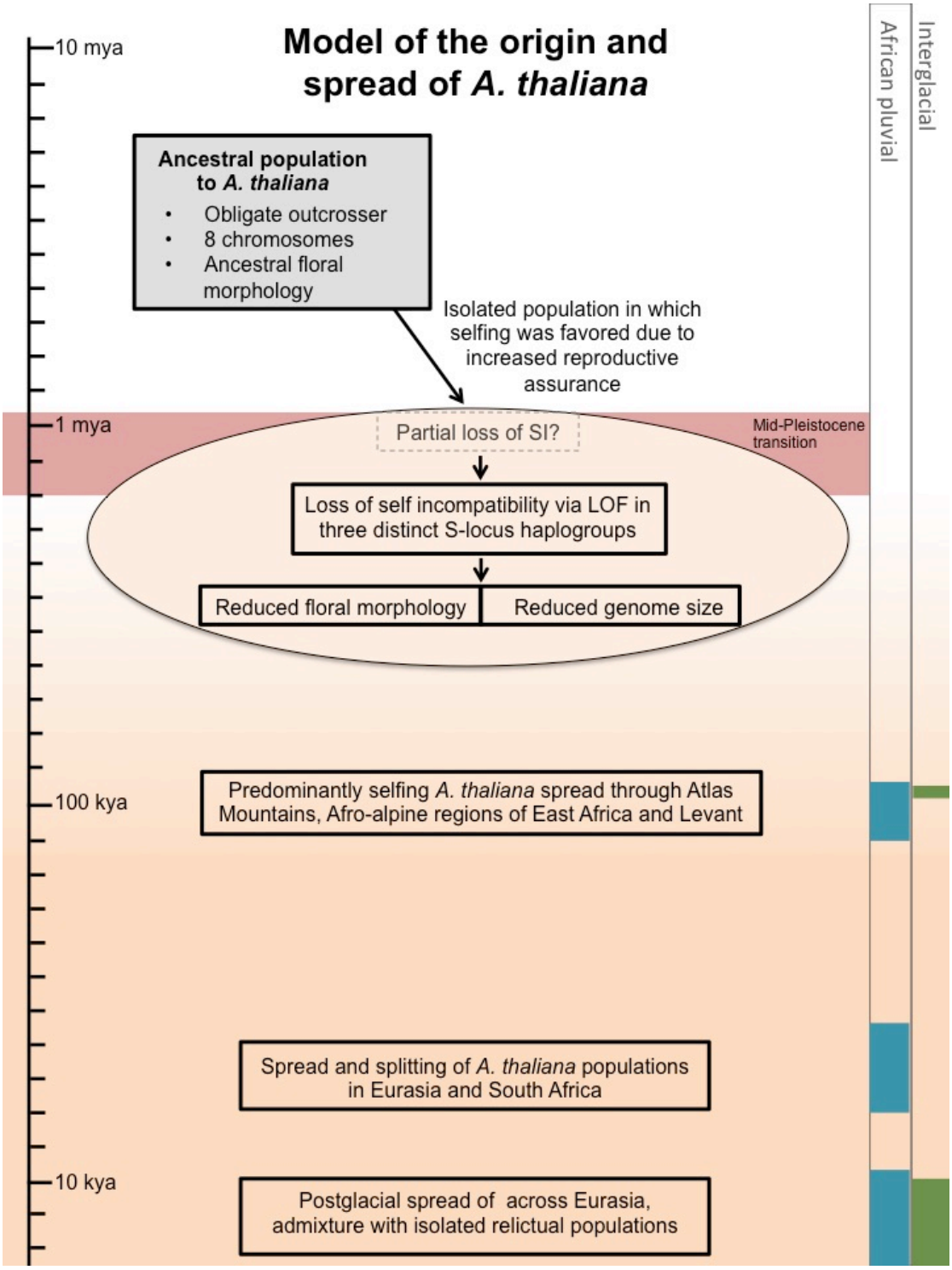


Figure S12. A graphical model of the origin of selfing and spread of *A. thaliana*. Inferred events are shown in boxes. Red bar denotes the Middle Pleistocene Transition (~1.2-0.8 mya) and pluvials and interglacials over the past 100 kya are shown in bars on the right.

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

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

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

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

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

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

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

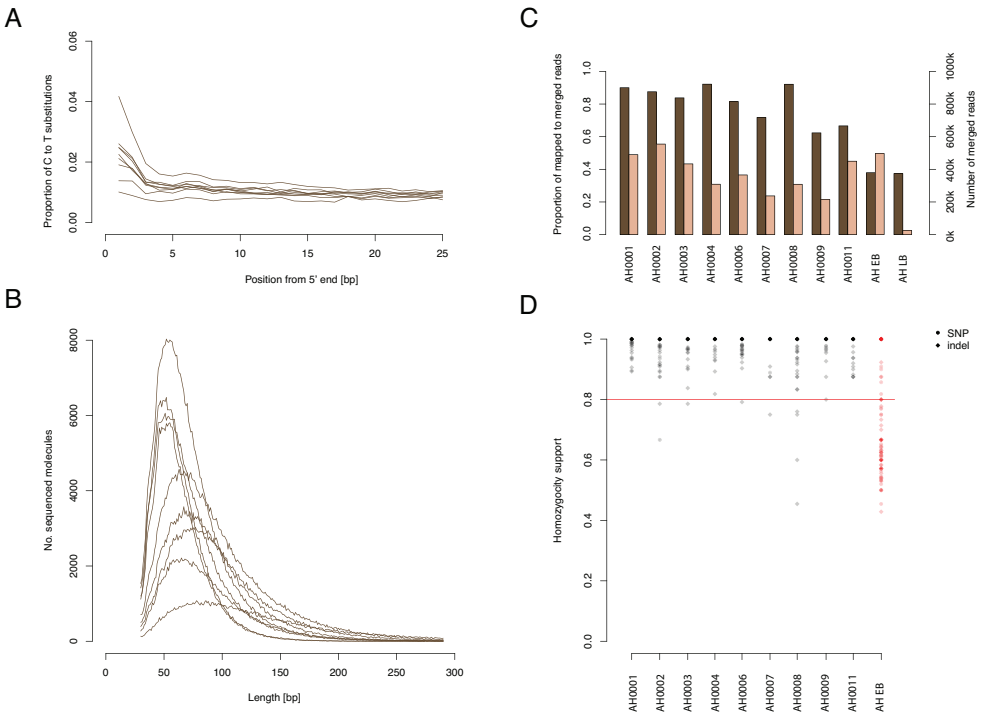


Figure S13. Authentication and analysis of herbarium ancient DNA (aDNA) samples.
A) Proportion of C-T transitions at the 5'-end of the reads. B) Read length distribution. C) Fraction of *A. thaliana* DNA as a proportion of mapped reads to merged reads (brown bars), and total number of merged reads (beige 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.

SUPPLEMENTARY TABLES

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. 'Cov[1]' shows Golden Path coverage and 'Cov[2]' 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-Kilamanjaro	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. 2*C*. 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	t	108.55	150.32	159.03	114.56	229.30
		d.f.	527.82	628.09	642.70	707.61	534.56
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Iberia, Balkans, Caucasus	t	88.42	101.52	106.38	59.19	183.78
		d.f.	601.45	884.43	911.10	983.93	624.11
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Northern Sweden	t	109.13	150.92	159.56	115.36	233.09
		d.f.	530.79	640.88	656.80	726.87	516.73
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Southern Sweden	t	105.41	143.79	152.15	106.46	221.38
		d.f.	528.59	631.41	646.36	712.64	529.71
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Central Europe	t	95.95	121.03	127.61	79.61	202.44
		d.f.	553.19	732.10	755.87	850.95	549.82
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Germany	t	112.62	163.31	173.56	130.78	238.44
		d.f.	506.78	534.39	538.55	557.78	506.21
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Western Europe	t	104.85	141.34	149.34	103.66	219.73
		d.f.	535.87	662.44	680.46	758.41	540.33
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Iberian non-relicts	t	83.30	99.55	105.85	50.99	174.02
		d.f.	522.68	605.60	617.82	672.90	549.51
		p	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16	<2.2E-16
	Iberian relicts	t	19.63	-10.38	-9.92	-43.30	70.97
		d.f.	960.68	794.19	768.27	688.03	925.26
		p	<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.