Supplementary Appendix

Demography and mating system shape the genome-wide impact of purifying selection in *Arabis alpina*

Benjamin Laenen[a,1], Andrew Tedder[a,1], Michael D. Nowak[a], Per Toräng[b], Jörg Wunder[c], Stefan Wötzel[c], Kim A. Steige[a,d], Yiannis Kourmpetis[e,f], Thomas Odong[e], Andreas D. Drouzas[g], Marco Bink[e,h], Jon Ågren[b,2], George Coupland[c,2], Tanja Slotte[a,2]

[a]Dept. of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University,
[b]Dept. of Ecology and Genetics, EBC, Uppsala University
[c]Max Planck Institute for Plant Breeding Research, Cologne
[d]Current address: Institute of Botany, Biozentrum, University of Cologne, 50674 Cologne, Germany
[e]Biometris, Wageningen University & Research Centre, P.O. Box 16, 6700 AC Wageningen, The Netherlands
[f]Current address: Nestlé Institute of Health Sciences, EPFL Campus, 1015 Lausanne, Switzerland
[g]School of Biology, Aristotle University of Thessaloniki
[h]Current address: Hendrix Genetics Research, Technology & Services B.V., P.O. Box 114, 5830 AC Boxmeer, The Netherlands

[1]B.L. and A.T. contributed equally to this work.
[2]Authors for correspondence, email: Jon Ågren: jon.agren@ebc.uu.se; George Coupland: coupland@mpipz.mpg.de, Tanja Slotte: Tanja.Slotte@su.se

Supplementary Text

Supplementary Methods

Data and sequencing

We sampled 38 *Arabis alpina* individuals throughout its European range (Table S1). DNA was extracted from leaf material using either a Qiagen DNeasy plant mini kit (Qiagen, Inc., Valencia, CA, USA) or a modified cetyl trimethyl ammonium bromide (CTAB) extraction method. Whole genome libraries with an insert size of 300-400 bp were prepared using the TruSeq DNA v2 protocol. Sequencing of 100bp paired-end reads was performed on an Illumina HiSeq 2000 instrument (Illumina, San Diego, CA, USA). The sequencing resulted in a total of 10,079 Gbp and ~245 Gbp (QC > 30) per sample on average with a mean coverage of 26X ranging from 16X to 45X.

Quality assessment, trimming and genotype calling

Sequencing adapters were identified using cutadapt v.1.8 (1) and trimmed from the raw sequences using Trimmomatic v.0.32 (2). Trimmed paired-end sequence reads and singleton reads whose pairs were removed in the trimming process were each mapped to the *A. alpina* V4 reference genome assembly using BWA-MEM v0.7.8 (3). Duplicated reads resulting from PCR replicates were removed using MarkDuplicates in Picard v2.0.1 (http://broadinstitute.github.io/picard/) and the resulting BAM alignment files were processed using the Genome Analysis Toolkit (4). Indels were realigned using GATK RealignerTargetCreator and IndelRealigner, and base quality scores were recalibrated using GATK BaseRecalibrator and PrintReads, all using default parameters. SNPs and indels were called separately with GATK UnifiedGenotyper using the DISCOVERY genotyping mode and default parameters. Genotype calling resulted in 13,410, 613 bi-allelic SNPs prior to filtering.

Filtering

Using GATK's SelectVariants and VariantFiltration modules, we selected SNPs which passed the following hard filters: quality-by-depth (QD) > 2.0; mapping quality (MQ) > 40.0; strand bias (FS) < 60.0; mapping quality rank sum test (MQRankSum) > -12.5; or a rank sum test (ReadPositionRankSum) > -8.0. Any variant site containing more than two alleles was removed from the data set, along with variant sites with less than 10X and more than 100X coverage.

With the aim of identifying problematic variant sites, we called SNPs in the *A. alpina* V4 genome assembly using the original shotgun sequence data of Willing et al. (5) using the procedures outlined above. The *A. alpina* accession sequenced by Willing et al. (5) was the result of five generations of self-fertilization with single seed descent and thus we expect there to be virtually no heterozygosity in this individual. Heterozygous sites identified in resequencing data from this individual were therefore removed from our data set. Furthermore, 20 kb windows that contained 10 or more reference heterozygous sites were removed.

The *A. alpina* genome assembly is exceptionally enriched for repetitive elements relative to previously sequenced Brassicaceae relatives (5). To avoid inflating our variant calls with reads mapped to repetitive elements, we removed all sites that fall within regions of the genome annotated as complex or simple repeats (e.g. TEs and microsatellites). Complex repeat annotations were taken directly from Willing et al.

(5), but simple repeats (mono-, di-, and tri-nucleotide repeats) were annotated using RepeatMasker v.4.0.1 (6). We also plotted the cumulative distribution of the proportion of 20 kb windows annotated as repeats along the genome, and based on this we chose to remove any 20 kb window with greater than 50% repeat annotation (i.e. ~6000 windows).

Sites with fixed heterozygosity across the whole dataset, which likely represent erroneous mapping in repeat regions and thus incorrect SNP calls were removed, along with sites with more than twenty percent missing data. Only sites that were anchored to the 8 pseudo-chromosomes in the assembly were kept (234 Mb).

To avoid calling SNPs in regions of the genome that likely represent copy number variants, we calculated average coverage in 20 kb windows for each sample and removed windows with coverage higher than 165X based on the cumulative distribution per sample, removing 687 windows in total. After subsetting the dataset into regional populations (see below) we noticed that 80% of the SNPs in the highly selfing Scandinavian population showed fixed heterozygosity indicative of incorrect SNP calls, likely due to repeat variants not caught by the previous filters. We used a sliding window (5kb window size, 1 kb step length) approach per individual to design a custom filter to detect regions with higher than average coverage. Specifically, we calculated both a coverage ratio (median coverage per 5kb window divided by the genome wide median coverage) and this ratio for the upper 95% percentile coverage per window to avoid regions with very high coverage. Per window ratio was averaged across all individuals, and windows with a ratio above a fixed threshold were excluded. We tested six median coverage ratio thresholds (1.1, 1.2, 1.3, 1.4, 1.5 and 2) and used 4 as a threshold for the upper 95% percentiles. In practice, thresholds of 2 and 4 for an individual with a median coverage of 30X would remove windows with a median coverage higher than 60X and windows with 5% of the sites having coverage higher than 120X. A median coverage ratio threshold of 2 and 4 as a threshold for the upper 95% percentile performed best, resulting in the removal of 87% of the fixed heterozygous sites from the Scandinavian population while maintaining 72% of the total dataset. After applying all filters the dataset was composed of 1,514,615 SNPs and 43,209,020 filtered invariant sites amenable for further analysis (Table S2).

Inference of population structure

We used 25,505 4-fold synonymous SNPs, pruned for linkage disequilibrium (LD) using PLINK v1.9 (7), to infer the population genetic structure of our *A. alpina* samples. We performed principal component analysis (PCA) using PLINK v1.9, and Bayesian clustering analysis (Fig. 1) using both fastSTRUCTURE v1.0 (8) and TESS v3 (9). In both cases, we tested values of *K* ranging between 2-20, with three replicate runs per *K*. For fastSTRUCTURE, optimal *K* was chosen using a combination of the ‘chooseK’ script, and cross validation error, and for TESS v3, cross entropy scores were used to determine optimal *K* value. Geographic maps of ancestry coefficients were generated using POPSutilitites.R (http://membres-timc.imag.fr/Olivier.Francois/pops.html) in R. v 3.2.3 (R Core Team, 2015), and individual ancestry coefficients were plotted using pophelper v1.1.6 (10) in R.

Pairwise F_{ST} estimates were obtained using the Weir and Cockerham (1984) estimator (11), as implemented in VCFTools 0.1.15 (12). We used 74,529 4-fold synonymous SNPs and estimated F_{ST} in 500 kb non-overapping windows for each pairwise comparison between regional populations. We report mean and 95% confidence intervals for each comparison.

Runs of homozygosity and decay of linkage disequilibrium

Progeny-array based outcrossing estimates have shown that populations from Scandinavia are highly selfing (up to ~10% outcrossing) whereas intermediate outcrossing rates have been estimated for two French and Spanish populations (~20% and ~ 18%, respectively) (13). We estimated runs of homozygosity to assess whether genomic data supported progeny-array based estimates of mating system variation. Following the recommendations of Howrigan, Simonson and Keller (14) we pruned the whole dataset for moderate linkage disequilibrium (LD) removing all SNPs within a 50 SNP window which showed an $r^2 > 0.5$ using PLINK v1.9 (7). For each of the five regional populations we performed a search for runs of homozygosity (ROH) in 100kb windows. A ROH was defined as an unbroken run of a minimum of 35 homozygous SNPs. ROH were binned into length categories, small (100-200kb), medium (200kb -500kb) and large (>500kb) (Supplementary Fig. S1).

We used popLDdecay (https://github.com/BGI-shenzhen/PopLDdecay) to estimate the decay in linkage disequilibrium (LD), using all available SNPs for each regional population (Table S4). We estimated the r^2 statistic using default parameters, which include a maximum distance of 300 bp between two SNPs, a minimum allele frequency of 0.005. Using 200 bp non-overlapping windows, we estimated the mean and 95% confidence interval for r^2 (Supplementary Fig. S2).

Summary statistics and inference of selection

Each of the five regional populations defined by fastSTRUCTURE was refiltered for missing data. Per population summary statistics (S, π, Tajima's D) were calculated for 4-fold degenerate sites and 0-fold degenerate sites, as described in (15). We also obtained estimates for introns and intergenic regions with low gene density and high recombination rate. These intergenic regions were selected to be less affected by linked selection, such that they would be useful for demographic inference, and had lower than the median gene density and higher than the third quartile of recombination rates. We further obtained population genetic summary statistics for total sites using global estimates and along fixed windows of 20kb (Table 1 and Table S4).

We used DFE-alpha v. 2.15 (16) to estimate the distribution of fitness effects (DFE) for new 0-fold degenerate nonsynonymous mutations. These analyses were based on folded 4-fold and 0-fold site frequency spectra for each population, with 4-fold degenerate synonymous sites assumed to be evolving neutrally. DFE was estimated for each population separately under a model with stepwise change in population size between two epochs as implemented as a built-in procedure in DFE-alpha, and each population's DFE was summarized in three bins representing increasing purifying selection ($0<N_e s<1$; $1<N_e s<10$; $N_e s>10$). Confidence intervals were generated with 200 bootstrap replicates, resampled using 10 kb windows restricted within chromosomes (Fig 2A, Table S7). Pairwise comparisons of each bin of the DFE were conducted, with FDR correction of the resulting P-values.

Major effect mutations and genetic load

We characterized the presence and frequency of major effect mutations in each population using snpEFF v4.2 (17). We focused on loss of start and stop codons, gain of stop codons and changes in splice sites. To avoid reference-biased inference of the alternate alleles we used an outgroup to polarize SNPs. We made a whole genome alignment of *A. alpina* V4 *and A. montbretiana* (assembly ASM148412v1; 5) using LASTZ v1.02.00 (18) following Steige et al. (15). For each population, we ran snpEFF using the polarized reference and recorded fixed major effect mutation in a homozygous state within ROH. We counted the number of homozygous major effect and nonsynonymous derived homozygous genotypes as a proxy for the recessive genetic load, and the average number of derived major-effect or nonsynonymous alleles per individuals as a proxy for the additive genetic load (see e.g.(19) (Table S5). For each population we also counted the number of fixed derived nonsynonymous or major-effect alleles after removing all missing data to have a constant SNP set among populations (Table S5). In order to test if genetic load show a similar pattern at highly conserved sites, we used Phastcons scores (20) from (21). We aligned *A. alpina* V4 to *A. lyrata* (ssp. *lyrata*; PRJNA41137) to assign a phastcons score to *A. alpina* sites that were aligned. We estimated the recessive and the additive genetic load for nonsynonymous sites with a Phastcons score >0.9 representing highly conserved sites among the nine Brassicaceae species used in (21).

Demographic history

To estimate parameters associated with the origin of Scandinavian *A. alpina*, we inferred the parameters of three demographic models (Table S6, Fig. S5) in fastsimcoal2 v. 2.5.2.21 (22), using two-dimensional joint SFS (2D-SFS) based on a scattered sample from central Europe (13 individuals from France, Switzerland, Germany & Poland) and the Scandinavian population with the Icelandic sample (9 individuals; Fig. 1A). This 2D-SFS was derived from 12,967 intergenic SNPs in regions with low gene density and high recombination rates (Fig S5).

We used 100,000 simulations to estimate log-likelihood, expected SFS, and a suite of model specific demographic parameters. To obtain global maximum likelihoods, we performed 50 independent replicate runs, with 10-40 conditional maximisation algorithm cycles and a mutation rate of 7 × 10-9 (23) and a generation time of 1.5 years to convert estimates into units of years and individuals. We based our choice of generation time on a population survey that estimated the lifetime expectancy in *A. alpina* populations to between 1.4 and 2.1 years (24). Confidence intervals were generated by performing parametric bootstrapping with 100 bootstrap replicates, and 50 runs per bootstrap. Model comparison was based on global maximum likelihood using the Akaike information criterion with a correction for finite sample sizes (AIC_C; 25) and Akaike's weight of evidence calculated using the qpcR v1.40 package in R. v 3.2.3. We assessed the fit of the best model by comparing the observed SFS and Tajima's D with results from 1000 coalescent simulations in fastsimcoal2 v. 2.5.2.21 (Fig S7).

We further estimated the demographic history for each regional population with StairwayPlot v0.2.beta (26) using SFSs for intergenic SNPs in regions with low gene density and high recombination rates. We report changes in effective population size (N_e) for the best fit model (Fig S8).

Forward simulations

We used forward simulations to assess the impact of demography and selection associated with a shift to selfing on genetic diversity in the Scandinavian population using SLiM2 v2.1 (27). We simulated data using real exon positions on the 8 chromosomes, constant mutation rates ($7*10^{-9}$ base substitutions per sites per generation; (23)) and a recombination rates map in 50kb windows derived from a RAD-seq linkage map (28) based on a cross of two French accessions. We used a distribution of fitness effects based on estimates for the Greek population (Table S7) where large effective population size and obligate outcrossing should improve the accuracy of the estimation. In order to test if the results were robust to a change in the DFE, we also simulated using the DFE for the Central European population (Fig S9). We simulated our data under two competing, two population demographic models based on the *fastsimcoal2* results. We first ran the simulation for 10,000 generation before the split, after which a shift to 90% selfing was implemented for the population representing Scandinavia. In model one, this population remained constant (N_e = 1000), whereas it was subjected to a ten-fold bottleneck (N_e = 100) in model two. The Central European population had a constant population size (N_e = 1000) in both models. Simulations were sampled at two times points (12000 ybp and 20208 ybp), which corresponds to macrofossil evidence (29) for the presence of *A. alpina* in Scandinavia and to inferred time of divergence between Central European population and Scandinavian population (Fig. 3A). Mutation rates and recombination rates were scaled to simulate genetic diversity close to the observed data. Thirteen samples were randomly drawn from the simulated Central European population and eight from the simulated Scandinavian population. Neutral diversity at 4-fold synonymous sites was recorded in each population at the two time points mentioned above. We note that in these simulations we used a linkage map based on accessions that were not from Scandinavia. It is possible that crossover rates could be higher in selfers (30), and if this were the case, we would expect to see a less pronounced effect of background selection on neutral diversity in selfing populations from Scandinavia. These simulations should thus be conservative with respect to assessing whether the reduction in diversity in Scandinavia can be explained by selfing and background selection alone, without a concomitant demographic change.

Supplementary Results

ROH and decay of LD support mating system variation inferred using progeny array estimates of outcrossing rates

Inbreeding increases homozygosity, resulting in longer blocks of contiguous homozygous genomic tracts, termed runs of homozygosity (ROH) (31). In the absence of additional confounding effects, we therefore expect lengths of ROHs to be highest for self-fertilizing populations, followed in turn by mixed-mating populations and outcrossing populations. We quantified ROH based on 474,250 SNPs pruned for LD, and found that in agreement with our expectation, highly self-fertilizing Scandinavian individuals have very long ROH, whereas mixed-mating French and Spanish individuals have intermediate and variable lengths of ROH, and outcrossing Greek individuals harbor few and typically short ROH (Fig. S1). Italian individuals show markedly longer ROH than Greek individuals. In good agreement with these results, LD decayed the fastest with physical distance in the Greek population, followed by the Italian, French and Spanish populations. In contrast, the Scandinavian populations exhibited high long-range LD (albeit with broad confidence intervals, likely as a result of the low number of SNPs available for analysis in this population) (Fig. S2). Given the evidence for functional self-incompatibility in the Italian population, the intermediate patterns of LD and ROH in this population suggests the action of additional factors that affect homozygosity beyond outcrossing rates, for instance biparental inbreeding or bottleneck events.

Supplementary Tables

Table S1. Geographical origin of all included samples and average coverage of resequencing data.

Sample ID	Country	Location	Latitude	Longitude	Coverage
AaVikS1	Greece	Vikos	39.9534	20.7043	32.6
AaVikS4	Greece	Vikos	39.9534	20.7043	27.5
AaVikS6	Greece	Vikos	39.9534	20.7043	27.7
AaVikS7	Greece	Vikos	39.9534	20.7043	26.5
AaVikS8	Greece	Vikos	39.9534	20.7043	27.0
523_I	Italy	Apuan Alps	44.1300	10.2100	26.7
Aa157-11A	Italy	Apuan Alps	44.0788	10.3270	21.2
Aa157-3A	Italy	Apuan Alps	44.0788	10.3270	22.9
Aa157-4A	Italy	Apuan Alps	44.0788	10.3270	21.7
Aa157-7A	Italy	Apuan Alps	44.0788	10.3270	21.6
222_A	France	Pic Blanc	45.0641	6.3839	18.4
222_B	France	Pic Blanc	45.0641	6.3839	45.1
222_C	France	Pic Blanc	45.0641	6.3839	17.1
222_D	France	Galibier	45.0605	6.4036	20.8
222_E	France	Galibier	45.0605	6.4036	16.8
222_F	France	Galibier	45.0605	6.4036	16.0
222_Q	France	Galibier	45.0605	6.4036	15.6
222_N	France	Granon	44.9663	6.5824	23.8
222_O	France	Granon	44.9663	6.5825	15.6
222_P	France	Granon	44.9663	6.5825	26.2
222_I	Spain	Lago de la Cueva	43.0515	-6.1048	28.6
222_J	Spain	Lago de la Cueva	43.0515	-6.1048	24.0
222_L	Spain	Angliru	43.2298	-5.9391	22.3
222_M	Spain	Angliru	43.2298	-5.9391	20.6
523_F	Spain	Jaca Pyrenees	42.5500	-0.5500	34.2
222_G	Sweden	Geargevaggi	68.4136	18.3197	28.8
222_H	Sweden	Geargevaggi	68.4136	18.3197	24.5
222_S	Sweden	Nuolja	68.3608	18.7169	25.5
222_T	Sweden	Nuolja	68.3608	18.7170	28.7
222_U	Norway	Riasten	62.8344	11.7445	27.6
222_V	Norway	Riasten	62.8344	11.7445	30.6
222_W	Sweden	Tväråklumpen	63.2082	12.3499	28.5
222_X	Sweden	Tväråklumpen	63.2082	12.3499	36.5
523_D	Iceland	Sveinstindur	64.1000	-18.3700	30.3
523_B	Poland	Czarna Gora	49.4283	20.1242	25.0

523_G	Portugal	Madeira	32.7608	-17.1342	24.8
523_A	Switzerland	Gornegrat	45.9836	7.7842	29.0
523_H	Germany	Wisent Tal	49.7907	11.2741	43.3

Table S2. Population genetic summary statistics and confidence intervals (CI) based on genomic resequencing of 38 samples of *A.alpina*. Estimates are shown for 0-fold degenerate nonsynonymous sites, 4-fold degenerate synonymous sites, intergenic sites in regions with low gene density and high recombination rates, and all sites.

Site type	**Invariant**	**SNPs**	**π**	**CI π**	**Tajima's D**	**CI Tajima's D**
0-fold	5991447	98564	0.0027	(0.0003 - 0.0089)	-0.70	(-2.00 - 1.18)
4-fold	1292547	65821	0.0102	(0.0007 - 0.0310)	-0.14	(-1.88 - 2.04)
intergenic high-recombination, low gene density regions	2578961	95844	0.0061	(0.0020 - 0.0137)	-0.71	(-1.96 - 0.90)
Total	43209020	1514615	0.0058	(0.0018 - 0.0129)	-0.66	(-1.92 - 0.90)

Table S3. Pairwise F_{ST} estimates for regional populations of *A. alpina*.

Comparison	**mean F_{ST}**	**median F_{ST}**	**2.5% CI bound**	**97.5% CI bound**
France vs Scandinavia	0.54	0.56	0.24	0.78
Greece vs France	0.42	0.42	0.16	0.69
Greece vs Italy	0.44	0.42	0.24	0.70
Greece vs Spain	0.46	0.46	0.21	0.73
Greece vs Scandinavia	0.73	0.75	0.52	0.89
Italy vs France	0.46	0.46	0.19	0.74
Italy vs Spain	0.49	0.48	0.23	0.76
Italy vs Scandinavia	0.82	0.83	0.66	0.95
Spain vs France	0.39	0.38	0.14	0.72
Spain vs Scandinavia	0.81	0.82	0.61	0.96

Table S4. Population genetic summary statistics (sd = standard deviation) for each regional population. Invariant sites, segregating sites (S), nucleotide diversity (π) and Tajima's D are reported.

Regional population (*n*)	**Site type**	**Invariant**	**S**	**π (sd)**	**Tajima's D (sd)**
Greece (5)	0-fold	6233152	41026	0.0022 (0.00214)	-0.35 (0.79)
	4-fold	1381058	30920	0.00803 (0.00761)	-0.09 (0.85)
	intron	11012738	180847	0.0056 (0.00379)	-0.26 (0.73)
	neutral interg.	1740608	27968	0.00545 (0.00325)	-0.28 (0.65)
	Total	47255711	669349	0.00481 (0.00275)	-0.24 (0.59)
Italy (5)	0-fold	6248282	23135	0.00151 (0.00162)	0.57 (0.88)
	4-fold	1392110	17971	0.00537 (0.00582)	0.59 (0.89)
	intron	11091311	99431	0.00369 (0.00286)	0.75 (0.83)
	interg. high rec. high gene density	1752304	14752	0.00339 (0.00234)	0.71 (0.69)
	Total	47533084	367729	0.00312 (0.00195)	0.8 (0.76)
Spain (5)	0-fold	6251401	21233	0.00135 (0.00151)	0.42 (0.73)
	4-fold	1394351	15852	0.00463 (0.00559)	0.49 (0.78)
	intron	11118215	88697	0.00323 (0.00286)	0.49 (0.68)
	interg. high rec. high gene density	1757114	13471	0.00301 (0.00222)	0.49 (0.63)
	Total	47627475	334827	0.00279 (0.00194)	0.52 (0.58)
France (10)	0-fold	6231797	24189	0.00144 (0.00162)	0.85 (0.95)
	4-fold	1378448	19661	0.00522 (0.00586)	0.83 (1.02)
	intron	11058404	107087	0.00363 (0.00288)	1.16 (0.93)
	interg. high rec. high gene density	1747360	15238	0.00324 (0.00254)	1.1 (0.86)

	Total	47375073	395597	0.00309 (0.00203)	1.21 (0.85)
Scandinavia (8)	0-fold	6271581	1151	0.00007 (0.00034)	0.07 (1.19)
	4-fold	1410199	589	0.00017 (0.00099)	0.22 (1.32)
	intron	11192785	3898	0.00011 (0.00037)	-0.14 (1.16)
	interg. high rec. high gene density	1767719	599	0.00013 (0.00046)	-0.07 (1.23)
	Total	47910901	15632	0.00012 (0.00038)	-0.1 (1.17)

Table S5. Number of invariant sites, segregating sites and sites fixed for the derived allele based on an outgroup (*A. montbretiana*) in each population for 4-fold synonymous, 0-fold-synonymous and major effect mutations. A. Counts after removing all missing data in any individual and in any population. B. Counts for the full data set.

A.

	Greece	Italy	Spain	France	Scandinavia
4-fold					
Fixed ancestral	8711	11513	12878	10673	13865
Segregating	8144	4706	4036	6197	136
Fixed derived	2956	3592	2897	2941	5810
0-fold					
Fixed ancestral	15770	19700	21468	19093	23231
Segregating	11290	6536	5686	7900	328
Fixed derived	4374	5198	4280	4441	7875
Major effect mutations					
Fixed ancestral	268	304	330	310	357
Segregating	128	90	68	91	10
Fixed derived	43	45	41	38	72

B.

	Greece	Italy	Spain	France	Scandinavia
4-fold					
Fixed ancestral	26197	30652	33374	12070	39754
Segregating	23964	12623	10771	7151	449
Fixed derived	8758	9488	7328	3348	16563
0-fold					
Fixed ancestral	43318	49144	52265	21848	61956
Segregating	30870	15946	14321	9013	879
Fixed derived	12104	12872	10407	5084	21339
Major effect mutations					
Fixed ancestral	593	650	699	357	789
Segregating	304	187	136	99	21
Fixed derived	79	84	84	46	170

Table S6. Estimated parameters and model fit for three demographic models of the split between Central European and Scandinavian *A. alpina*. The preferred model (Akaike weight 0.924) has a population split followed by a bottleneck, and no subsequent migration. The model with migration has unlimited bidirectional migration for 2000 generations following the population split. All estimated times are given in years before present (ybp), assuming a generation time of 1.5 years. Maximum likelihood estimates are shown, with 95% confidence intervals in parenthes.

Model	$N_{Central\ Europe}$[1]	$N_{Scandinavia}$[2]	N_{BOT}[3]	T_{DIV}[4]	T_{BOT}[5]	Mig_{CES}[6]	Mig_{SCE}[6]	AIC (w)[7]
Split + Bottleneck	128691 (122774 - 134202)	9660 (8685 - 11790)	9218 (6633 - 11115)	20208 (18999 - 24218)	2697 (509 - 22188)	NA	NA	98506 (0.924)
Split	136092 (128513 - 144295)	27771 (23932 - 29928)	NA	59288 (56993 - 65093)	NA	NA	NA	100387 (0)
Split + Bottleneck + migration	127391 (122316 - 134315)	10331 (8528 - 11109)	9887 (7094 - 10270)	52236 (45627 - 58140)	22108 (10591 - 54262)	3.13E-11 (2.23E-13 - 6.95E-7)	4.70E-13 (3.76E-14 - 3.10E-8)	98511 (0.076)

[1] Effective population size estimate for our scattered sample of Central European *A. alpina*.
[2] Effective population size estimate for Scandinavian *A. alpina*.
[3] Effective population size for duration of bottleneck.
[4] Population split time.
[5] Bottleneck end.
[6] Migration rate, Mig_{CES} corresponds to probability that a Central European individual originates from Scandinavia and Mig_{SCE} to the probability that a Scandinavian individual originates in Central Europe.
[7] Akaike information criterion, weight

Table S7. Results of DFE-alpha analysis of selection on 0-fold non-synonymous mutations. Analyses assumed a two epoch demographic model for each regional population. *b* and $N_e s$ correspond to the shape and mean of the gamma distribution used to model the DFE. The proportion of sites under increasing level of purifying selection is shown in bins of $N_e s$, from nearly neutral ($N_e s$ 0-1), through mildly deleterious ($N_e s$ 1-10) to highly deleterious ($N_e s$ > 10).

Regional population	***b***	**$N_e s$**	**$N_e s$ 0-1**	**$N_e s$ 1 - 10**	**$N_e s$ >10**
Greece	0.26 (0.21, 0.30)	-1.26E+2 (-4.06E+02, -8.76E+01)	0.22 (0.2, 0.24)	0.18 (0.15, 0.21)	0.60 (0.58, 0.62)
Italy	0.13 (0.05, 0.37)	-8.00E+3 (-2.62E+10, -5.85E+01)	0.25 (0.17, 0.28)	0.09 (0.03, 0.24)	0.66 (0.57, 0.71)
Spain	0.15 (0.06, 027)	-1.39E+3 (-7.83E+07, -9.44E+01)	0.24 (0.21, 0.28)	0.10 (0.04, 0.18)	0.66 (0.59, 0.68)
France	0.05 (0.05, 0.14)	-1.89E+10 (-5.23E+10, -7.20E+03)	0.27 (0.24, 0.28)	0.03 (0.03, 0.09)	0.69 (0.66, 0.71)
Scandinavia	0.42 (0.05, 99.99)	-1.24E+1 (-6.43E+06, -2.09E+00)	0.30 (0, 0.51)	0.40 (0.05, 1)	0.30 (0, 0.55)

Supplementary Figures

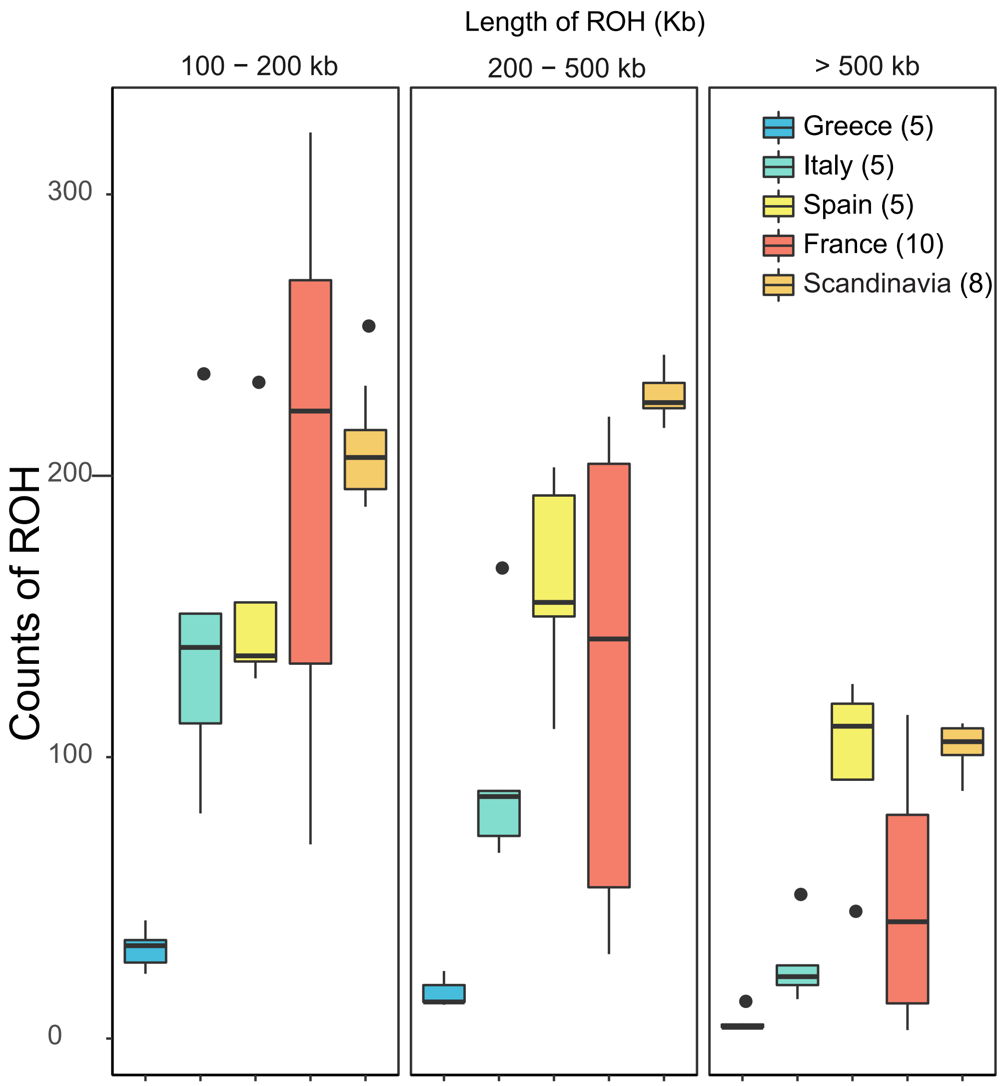

Figure S1. The number of runs of homozygosity (ROH) differs among regional populations with different mating systems. ROH were binned into small (100 – 200kb), medium (200 – 500kb) and large (>500kb) runs. Self-incompatible Greek individuals have the shortest ROH, followed by self-incompatible Italian individuals, mixed-mating Spanish individuals, and the longest ROH are found in highly self-fertilizing Scandinavian individuals. Individuals in the French cluster show highly variable lengths of ROH. Thick bars are the median count, box edges the interquartile and whiskers represent 1.5 times the interquartile range. For each regional population, the sample size is shown in parentheses in the figure legend.

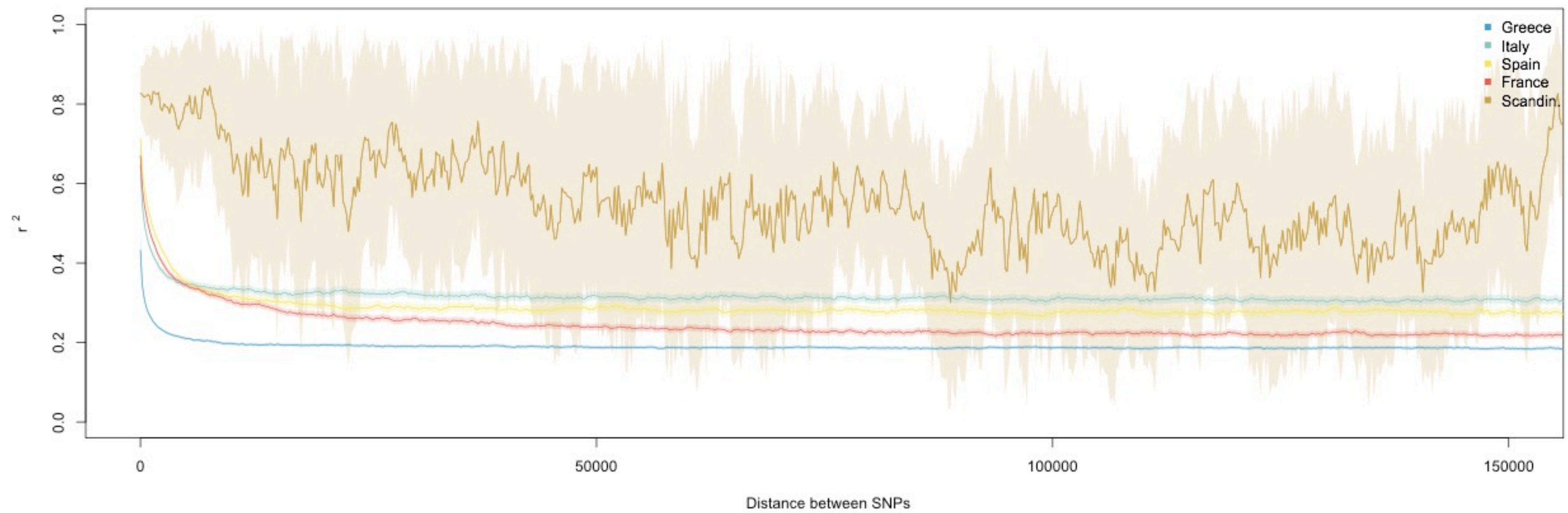

Figure S2. Decay of linkage disequilibrium (r^2) with physical distance for each regional population of *A. alpina.*

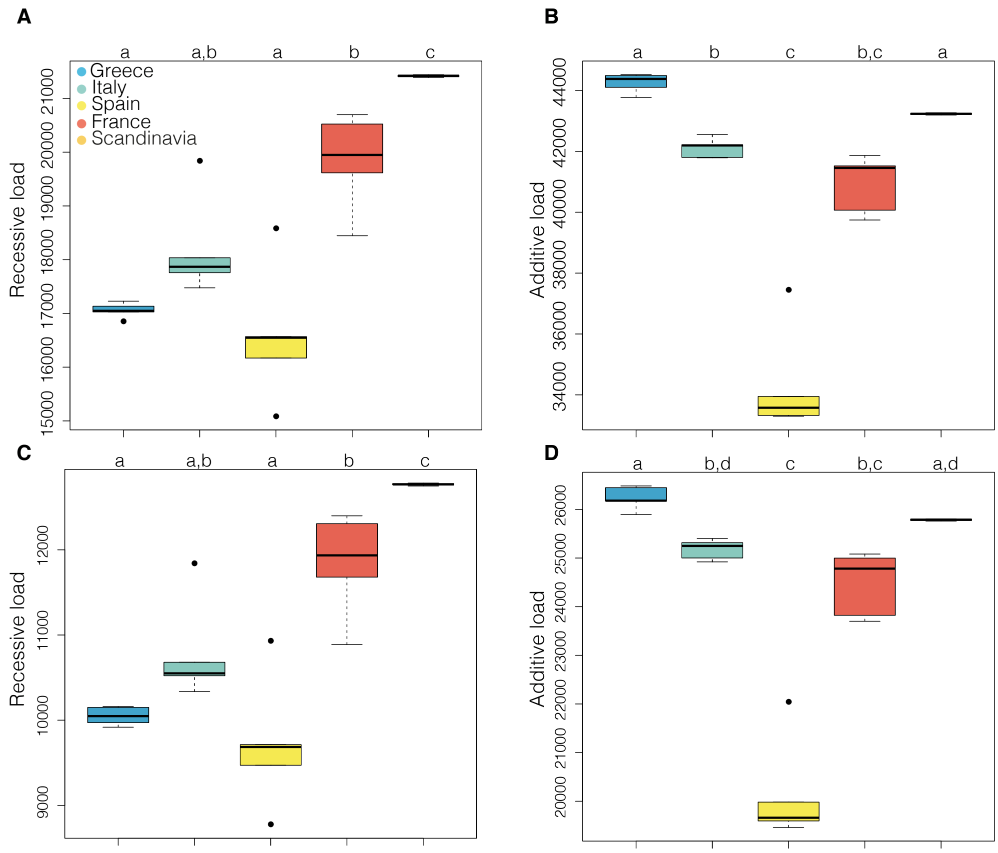

Figure S3. Genetic load estimates for outcrossing, mixed-mating and highly selfing *A. alpina* regional populations based on derived nonsynonymous alleles (A and B) and strongly constrained derived nonsynonymous alleles (C and D). Lowercase letters indicate groups with statistically significant differences (P<0.05) based on a Kruskal-Wallis test followed by post-hoc Dunn test. A. The recessive genetic load (number of derived homozygous genotypes) for 0-fold nonsynonymous variants. B. The additive genetic load (number of derived alleles) for 0-fold nonsynonymous variants. C. The recessive genetic load for 0-fold nonsynonymous variants at highly constrained sites, defined as those with Phastcons score >0.9 based on an analysis of nine Brassicaceae species. D. The additive genetic load for 0-fold nonsynonymous variants at highly constrained sites.

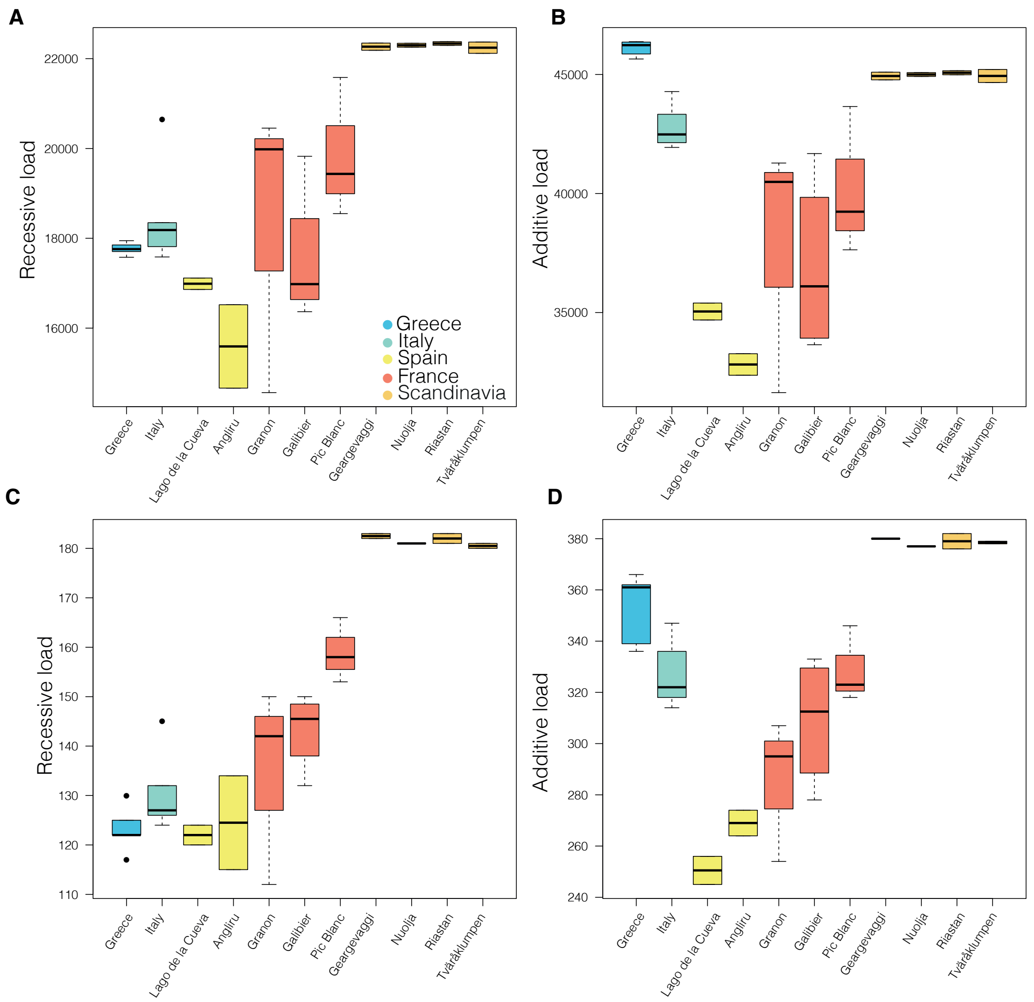

Figure S4. Genetic load estimates for outcrossing, mixed-mating and highly selfing *A. alpina* geographical populations based on derived 0-fold nonsynonymous alleles (A and B) and derived major-effect alleles (C and D). A and C show the recessive load (number of derived homozygous genotypes) whereas B and D show the additive genetic load (number of derived alleles).

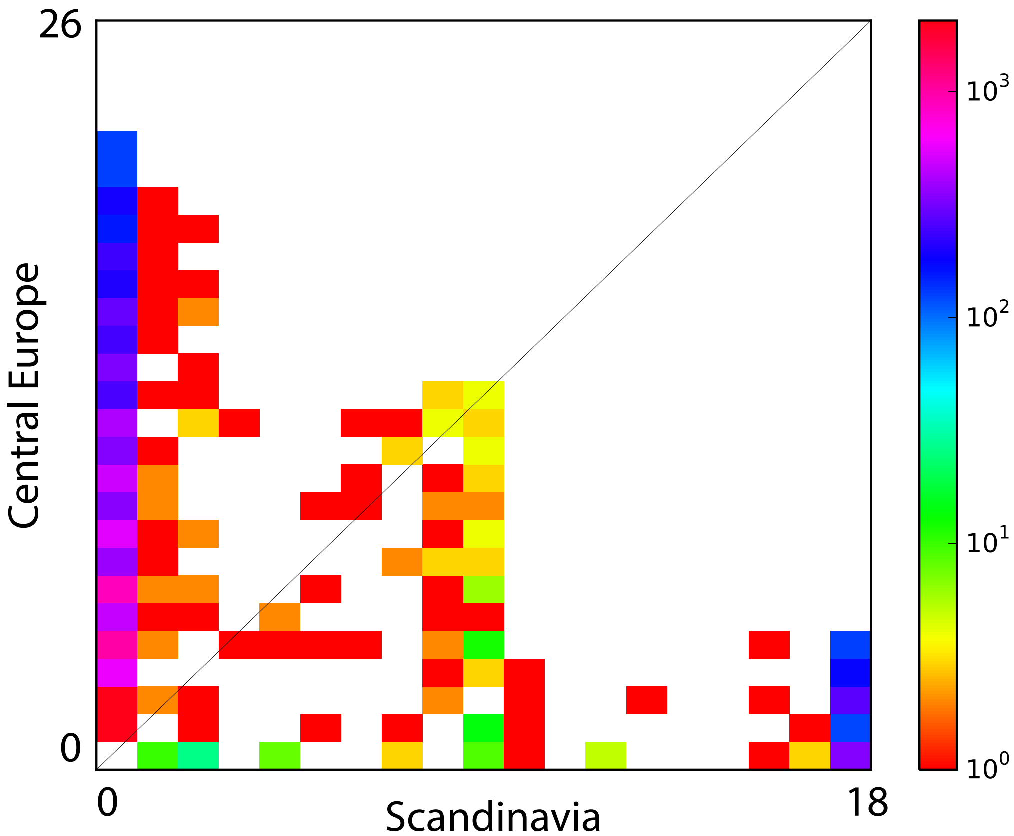

Figure S5. Joint site frequency spectrum for Scandinavian and Central European *A. alpina*.

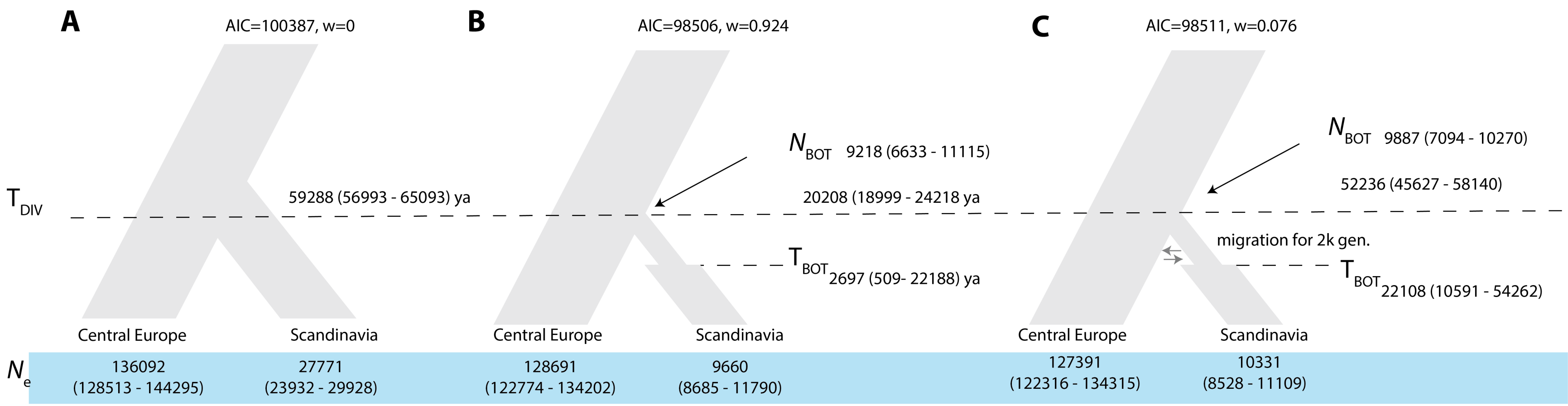

Figure S6. Three demographic models used to estimate divergence time and population size of the Scandinavian population of *A. alpina.* The model depicted in panel B is preferred based on AIC. Full parameter estimates are also shown in Table S6.

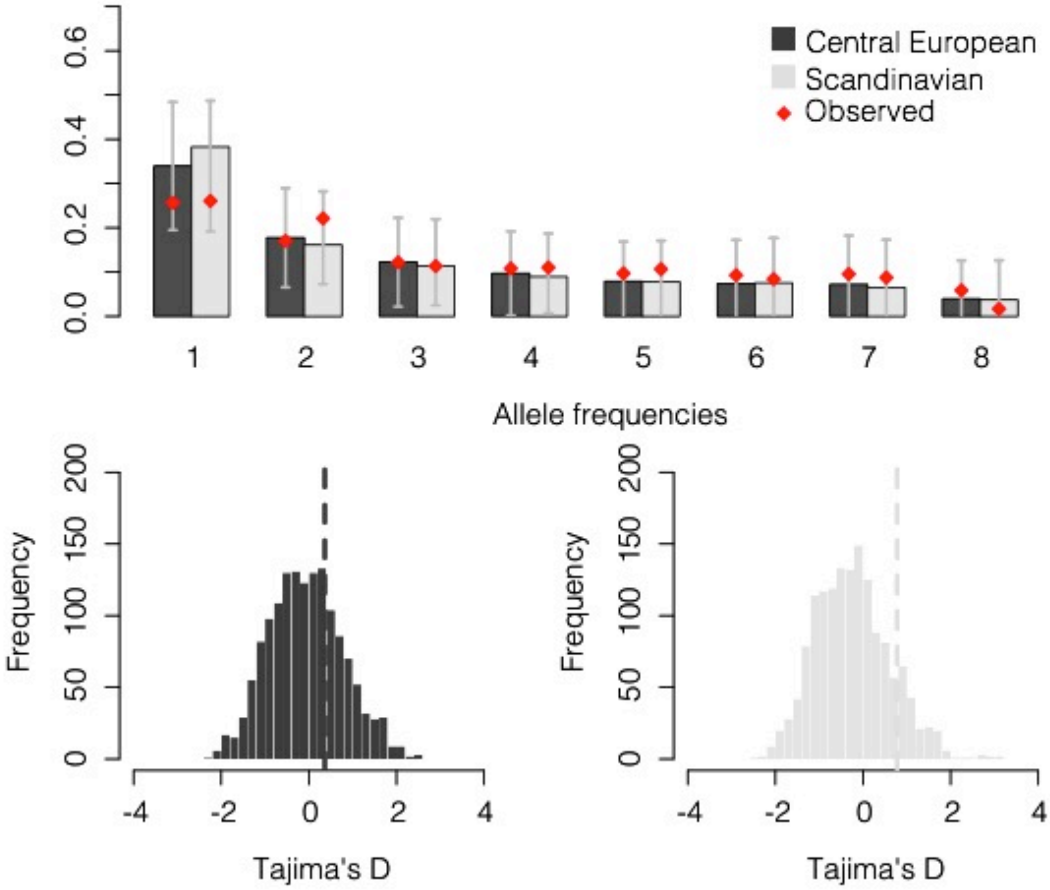

Figure S7. Folded site frequency spectra (SFS) and Tajima's D values derived from 1000 coalescent simulations under the best-fit demographic model compared to the observed SFS (red diamonds) and Tajima's D (dashed lines) for the Scandinavian and the Central European populations. Barplots and error bars correspond respectively to the average SFS and the standard deviation over the simulations. Note that the SFS have been downsampled to the same sample size to facilitate comparison.

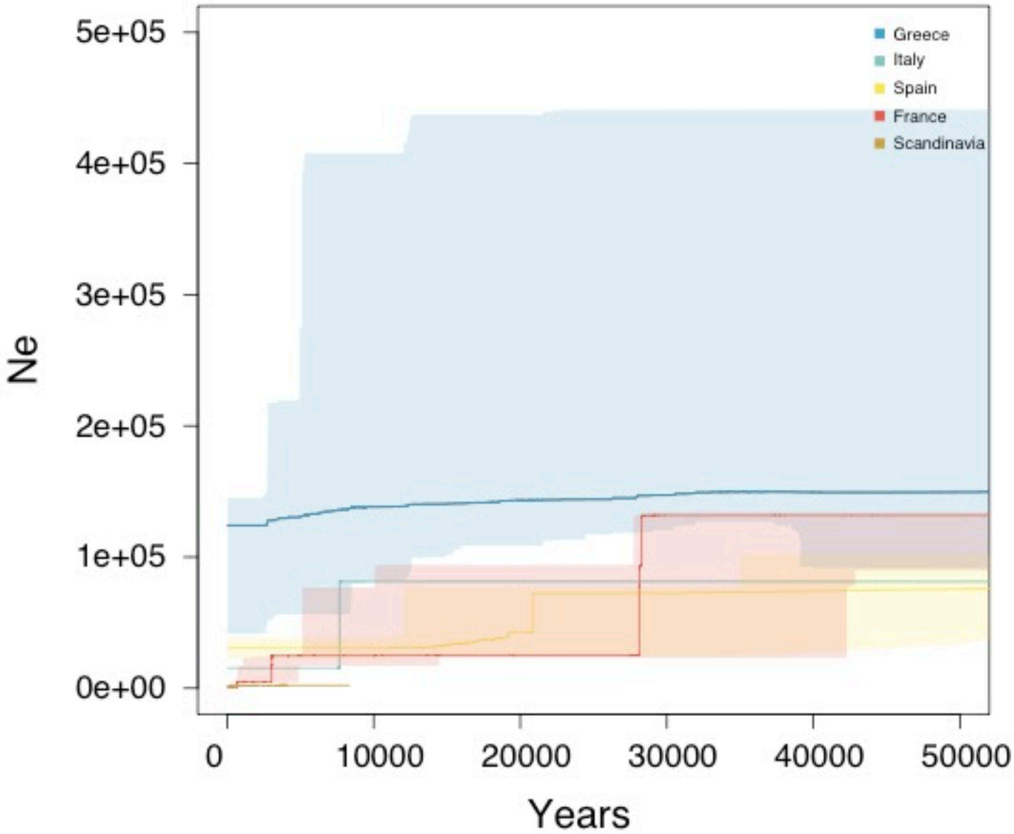

Figure S8. Recent population history inferred using stairway plot analyses. Lines correspond to best-fit estimates of N_e for each regional popuation whereas shaded areas indicate 95% confidence intervals.

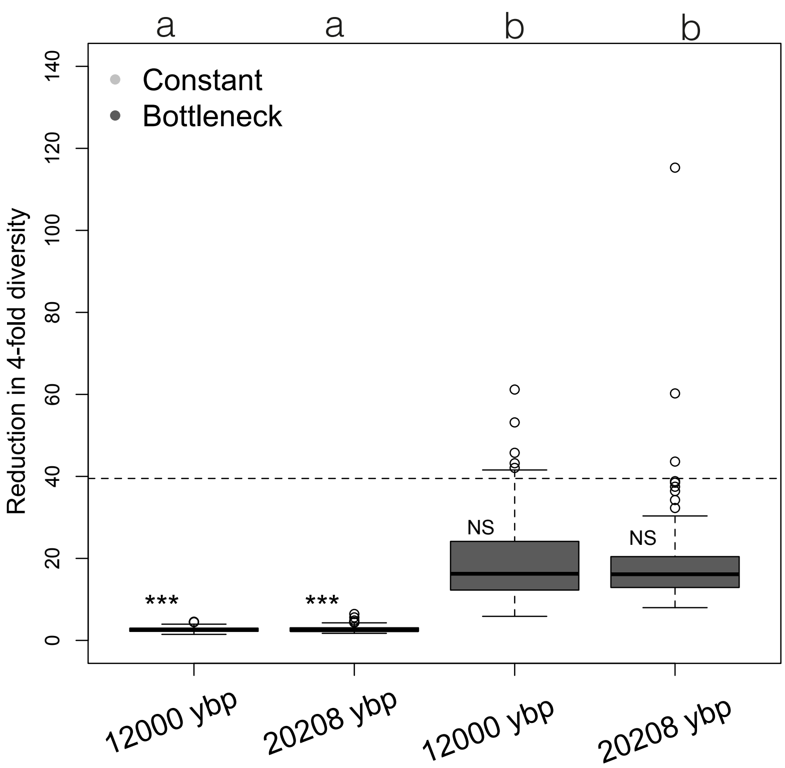

Figure S9. Results of forward simulations using the DFE derived from the Central European population. The boxplots show the ratio of synonymous polymorphism between an outcrossing population and a 90% selfing population experiencing either a constant population size or a 10-fold bottleneck, with the two populations diverging either 12,000 ybp or 20,208 ybp. The dashed line indicates the observed ratio of synonymous polymorphism in Central Europe to that in Scandinavia. Letters indicate significant difference between models (Mann-Whitney test $P<0.001$). Asterisks indicate an observed neutral diversity reduction significantly greater than that expected, based on 300 simulations.

References

1. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17(1):pp. 10-12.
2. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
3. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*.
4. 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.
5. Willing EM*, et al.* (2015) Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. *Nat Plants* 1:14023.
6. Smit AF, Hubley R, & Green P (1996) RepeatMasker Open-3.0.
7. 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.
8. Raj A, Stephens M, & Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197(2):573-589.
9. Caye K, Deist TM, Martins H, Michel O, & Francois O (2016) TESS3: fast inference of spatial population structure and genome scans for selection. *Mol Ecol Resour* 16(2):540-548.
10. Francis RM (2017) pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* 17(1):27-32.
11. Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 6: 1358-1370.
12. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.
13. Toräng, P., L. Vikström, J. Wunder, S. Wötzel, G. Coupland, and J. Ågren. 2017. Evolution of the selfing syndrome: Anther orientation and herkogamy together determine reproductive assurance in a self-compatible plant. *Evolution* 71:2206–2218..
14. Howrigan DP, Simonson MA, & Keller MC (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection
15. Steige KA, Laenen B, Reimegard J, Scofield DG, & Slotte T (2017) Genomic analysis reveals major determinants of *cis*-regulatory variation in *Capsella grandiflora*. *Proc Natl Acad Sci USA* 114(5):1087-1092.15.
16. Keightley PD & Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251-2261.
17. Cingolani P*, et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6(2):80-92.
18. Harris RS (2007) *Improved pairwise alignment of genomic DNA* (ProQuest).
19. Henn BM*, et al.* (2016) Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci USA* 113(4):E440-449.
20. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK,

Gibbs RA, Kent WJ, Miller W, Haussler D 2005 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. (2005) *Genome Res* 15(8):1034–1050.

21. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, Dewar K, Stichcombe JR, Schoen DJ, Wang X, Schmutz J, Town CD, Edger PP, Pires JC, Schumaker KS, Jarvis DE, Mandakova T, Lysak MA, van den Bergh E, Schranz ME, Harrison PM, Moses AM, Bureau TE, Wright SI, Blanchette M (2013) An atlas of over 90,000 conserved noncoding sequences provide insight into crucifer regulatory regions *Nat Genet* 45(8)891-898.
22. Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, & Foll M (2013) Robust demographic inference from genomic and SNP data. *Plos Genet* 9(10):e1003905.
23. Ossowski S*, et al.* (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* 327(5961):92-94.
24. Andrello M, de Villemereuil P, Busson D, Gaggiotti OE, Till-Bottraud I (2016) Population dynamics of *Arabis alpina* in the French Alps: evidence for demographic compensation? bioRxiv doi: https://doi.org/10.1101/070847
25. Johnson JB & Omland KS (2004) Model selection in ecology and evolution. *Trends Ecol Evol* 19(2):101-108.
26. Liu X, Fu Y-X (2015) Exploring population size changes using SNP frequency spectra. *Nat Genet* 47(5):555-559.
27. Haller BC & Messer PW (2017) SLiM 2: Flexible, Interactive Forward Genetic Simulations. *Mol Biol Evol* 34(1):230-240.
28. Jiao W-B, et al. (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res*. 27.5 (2017): 778-786.
29. Ehrich D*, et al.* (2007) Genetic consequences of Pleistocene range shifts: contrast between the Arctic, the Alps and the East African mountains. *Mol Ecol* 16(12):2542-2559.
30. Wright SI, Ness RW, Foxe JP, Barrett SCH. 2008. Genomic consequences of outcrossing and selfing in plants. *Int J Plant Sci* 169(1):105-118.
31. Gibson J, Morton NE, & Collins A (2006) Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* 15(5):789-795.