TITLE

Large haploblocks underlie rapid adaptation in an invasive weed

AUTHORS

Paul Battlay[1], Jonathan Wilson[1], Vanessa C. Bieker[2], Chris Lee[1], Diana Prapas[1], Bent Petersen[3,4], Sam Craig[1], Lotte van Boheemen[1], Romain Scalone[5,6], Nissanka P. de Silva[1], Amit Sharma[7], Bojan Konstantinović[8], Kristin A. Nurkowski[1,9], Loren Rieseberg[9], Tim Connallon[1], Michael D. Martin[2], Kathryn A. Hodgins[1]*

These authors contributed equally: Paul Battlay, Jonathan Wilson

These authors jointly supervised this work: Kathryn A. Hodgins, Michael D. Martin

[1] School of Biological Sciences, Monash University, Melbourne, Victoria, Australia

[2] Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[3] Center for Evolutionary Hologenomics, GLOBE Institute, University of Copenhagen, Denmark

[4] Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), AIMST University, 08100 Kedah, Malaysia

[5] Department of Crop Production Ecology, Uppsala Ecology Center, Swedish University of Agricultural Sciences, Uppsala, Sweden.

[6] Department of Grapevine Breeding, Hochschule Geisenheim University, Geisenheim, Germany.

[7] Cell, Molecular Biology and Genomics Group, Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[8] Department of Environmental and Plant Protection, University of Novi Sad, Novi Sad, Serbia

[9] Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, Canada

ABSTRACT

Adaptation is the central feature and leading explanation for the evolutionary diversification of life. Adaptation is also notoriously difficult to study in nature, owing to its complexity and logistically prohibitive timescale. We leverage extensive contemporary and historical collections of *Ambrosia artemisiifolia*—an aggressively invasive weed and primary cause of pollen-induced hayfever—to track the phenotypic and genetic causes of recent local adaptation across its native and invasive ranges in North America and Europe, respectively. Large haploblocks—indicative of chromosomal inversions—contain a disproportionate share of genomic regions conferring parallel adaptation to local climates between ranges (18%), are associated with rapidly adapting traits, and exhibit dramatic frequency shifts over space and time. These results highlight the importance of large-effect standing variants in rapid adaptation, which have been critical to *A. artemisiifolia*'s global spread across vast climatic gradients.

INTRODUCTION

Adaptation can be rapid, yet changes in the traits and genes that underlie adaptation are difficult to observe in real time because speed is relative: that which is fast in evolutionary terms is slow from the human perspective. Thus, while we know that adaptation is central to evolutionary diversification, species persistence, and invasiveness, the genetic and phenotypic dynamics of adaptation are difficult to document outside of the laboratory.

Invasive species are powerful systems for characterizing adaptation in nature, owing to several features that make them unique. In particular, biological invasions coincide with exceptionally rapid evolution[1–3], observable over human lifespans, as invasive populations can encounter drastically different environmental conditions from those of their source populations. Many invasions are, moreover, documented in large, geo-referenced herbarium collections, which can be phenotyped and sequenced to identify and track adaptive evolutionary changes through time—feats that are rarely achieved in natural populations. Invasive species frequently inhabit broad and climatically diverse ranges, which favours the evolution of adaptations to local environmental conditions[2,3], along with evolved tolerance of environmental extremes, which may be conducive to invasiveness[4]. Because they often occupy geographically broad native and invasive ranges, invasive species allow for tests of the predictability of evolution—a major puzzle in biology—as local adaptation across native and invasive ranges may favour either parallel or unique genetic solutions to shared environmental challenges. However, despite the promise of historical records and other features of invasive species that make them tractable systems for capturing adaptation in action, this treasure trove of data has not been fully utilized to elucidate the genetic basis of local adaptation during recent range expansions.

Ambrosia artemisiifolia, an annual weed native to North America, has mounted successful invasions on all continents except Antarctica[5]. The species thrives in disturbed habitats and has experienced extensive range shifts, historically documented in pollen records and herbarium collections. It also produces highly allergenic pollen, which is the chief cause of seasonal allergic rhinitis and asthma in the United States[6]. In Europe alone, approximately 13.5 million people suffer from *Ambrosia*-induced allergies, costing ~7.4 billion euros annually[7]. Continued invasion and climate change are predicted to more than double sensitization to *Ambrosia* pollen[8], further magnifying the health burden of this pest. Pollen monitoring has demonstrated that climate change has already significantly lengthened the ragweed pollen season, particularly at high latitudes[9]. Consequently, there is considerable incentive to understand the factors that

contribute to *Ambrosia* pollen production, including the species' invasive spread, the timing of pollen production, plant size, and fecundity.

Ambrosia artemisiifolia populations are characterized by strong local adaptation and high gene flow between populations[10,11]. In Europe, invasive populations have been established through multiple genetically diverse introductions from North America over the last ~160 years[11–13]. Remarkably, latitudinal clines observed for multiple traits in the native range, including flowering time and size, have re-evolved in Europe and Australia[14], suggesting rapid local adaptation following invasion. Moreover, this trait-level parallelism is echoed in signals of parallelism at the genetic level[10].

As biological invasions continue to increase worldwide[15] and the effects of anthropogenic climate change intensify, understanding the genetic architecture of adaptation to sudden environmental shifts—a classical question in evolutionary research—becomes ever more important. While long-standing theory suggests that evolution in response to incremental environmental change should proceed through mutations of infinitesimally small[16] or moderate[17] effect, large-effect mutations are predicted to be useful in bridging extreme, sudden environmental shifts[18]. Moreover, alleles of large effect will, in cases of local adaptation, be better able to persist in the face of the swamping effects of gene flow[19]. Large-effect mutations are also more likely to produce patterns of evolutionary repeatability, or genetic parallelism, between species' ranges[20], particularly if adaptive responses make use of standing genetic variation (as would be expected during a bout of rapid adaptation), rather than *de novo* mutations[21]. These features of large-effect mutations may, however, be achieved by groups of mutations in tight genetic linkage[19], including mutations captured by chromosomal inversions[22]. There is substantial empirical evidence for the involvement of inversions in local adaptation[23]. For example, *Drosophila melanogaster*'s *In(3R)Payne* inversion shows parallel environmental associations across multiple continents[24], and several plant inversions have been identified as contributing to local adaptation and ecotype divergence[25,26]. Theory also predicts that inversions can drive range expansions[27], though their actual contributions to biological invasions are not well-understood.

Here we develop a new, high quality and near chromosome-level reference assembly, and examine genome-wide variation in over 600 modern and historic *A. artemisiifolia* samples from throughout North America and Europe[11]. Using this data of unparalleled spatial and temporal

108 resolution, we first identified regions of the genome experiencing climate-mediated selection in
109 the native North American and introduced European ranges leveraging landscape genomic
110 approaches and genome-wide association studies of adapting traits such as flowering time.
111 Second, motivated by evidence that European and North American populations show similar
112 trait clines with respect to climate[14], we examined the extent of between-range parallelism at the
113 genetic level. Although adapting traits such as flowering time and size are polygenic, we
114 expected to see substantial levels of parallelism if large and moderate effect standing variants
115 were contributing to adaptive divergence. Third, we determined if these same regions show
116 temporal signatures of selection in Europe, which would be expected if some invading
117 populations were initially maladapted to their local climates. We coupled this temporal genomic
118 analysis with a temporal analysis of phenological trait changes in European herbarium samples
119 to further support our findings of genomic signatures of selection on flower time genes. Finally,
120 we identified haplotype blocks with multiple features consistent with inversions, in genomic
121 regions enriched for signatures of parallel adaptation. To determine if these putative structural
122 variants were contributing to rapid local adaptation in Europe, we assessed spatial and temporal
123 changes in their frequency as well as their associations with adapting traits.

124 # RESULTS

125 **Reference genome assembly**

126 We assembled a new, near chromosome-level *Ambrosia artemisiifolia* reference genome from a
127 heterozygous, diploid individual collected from Novi Sad, Serbia. After purging uncollapsed
128 heterozygosity[28] and scaffolding with HiRise[29] our final assembly consisted of 395 scaffolds,
129 with an L_{90} of 20 (*A. artemisiifolia* has 18 chromosomes[30]) and a genome size of 1.2Gb (flow
130 cytometry estimates of genome size range from 1.13-1.16Gb[31,32]). A total of 253 complete
131 Benchmarking Universal Single-Copy Orthologs (BUSCO[33]) genes (99.2%; 138 single copy
132 [54.1%]; 115 duplicated [45.1%]), no fragmented BUSCO genes and two missing BUSCO
133 genes (0.8%; table S1) were identified. The high numbers of duplicated BUSCO genes likely
134 reflect the whole genome duplications experienced in the Asteraceae, including a recent one
135 shared by *Helianthus* (sunflower) species at the base of the tribe[34,35]. This species also retained
136 a large number of duplicated BUSCO genes[35]. A large fraction of the genome consisted of
137 repetitive sequence (67%; table S2). Retroelements were the largest class (39.5%), with LTRs,
138 particularly Gypsy (7.87 %) and Copia (18.98 %), the most prevalent retroelements. MAKER [36]
139 identified 29,849 gene models with strong protein or transcript support, with average coding
140 lengths of 1.09kb, and 7.18 exons per gene (table S3; fig S1).

141

142 **Genome-wide association studies**

143 Genome-wide association studies (GWAS) using 121 modern samples across North American
144 (*n* = 43) and European (*n* = 78; table S4) ranges identified significant associations with 13 of 30
145 phenotypes, many of which are putatively adaptive, previously measured by van Boheemen,
146 Atwater and Hodgins[14] (fig. S2; table S5). Six phenotypes yielded associations within predicted
147 genes, including an association between flowering time phenotypes and a nonsynonymous SNP
148 in the *A. artemisiifolia* homologue of *A. thaliana* flowering-time pathway gene *early flowering 3*
149 (*ELF3*[37]), an "evolutionary hotspot" for parallel flowering time adaptation in *A. thaliana*, barley
150 and rice[38] (table S5; fig. 1A;B). Candidate SNPs in *ELF3* are restricted to high-latitude
151 populations in both ranges, where they occur at moderate to high frequencies (fig. 1C). While
152 the latitudes of these populations are greater in Europe than North America, the climatic
153 conditions are similar (fig. S3), indicative of local climate adaptation in parallel between ranges.

154

155 **Environmental-allele associations**

156 To identify genome-wide spatial signatures of local adaptation in North American and European
157 ranges, we performed genome scans for population allele frequencies among *A. artemisiifolia*

158 modern samples that were both highly divergent between populations (BayPass XtX)[39], and
159 correlated with environmental variables (latitude, longitude and 19 WorldClim temperature and
160 precipitation variables; table S6)[40]. Statistics were analyzed in 10kb windows using the
161 Weighted-Z Analysis (WZA)[41]. In North America (143 samples; 43 populations), 1,480 (84.0%)
162 of the 1,762 outlier windows for genomic divergence (XtX) were also outlier windows for at least
163 one environmental variable (XtX-EAA), while in Europe (141 samples; 31 populations), only 908
164 (51.7%) of the 1,755 XtX outlier windows overlapped environmental variable outlier windows.
165 Signatures of local adaptation were much stronger in North America than Europe, with the North
166 American range showing more extreme XtX values (fig. 1C), as well as more XtX-EAA windows
167 (table S6; fig. 1C). This suggests that North America exhibits greater population differentiation,
168 and a stronger relationship between population differentiation and the environment than Europe,
169 which is consistent with the expectation that samples from the native range will be better-
170 adapted to their environment than those from the recently-invaded European range.
171
172 Previous studies in *A. artemisiifolia* have identified signatures of repeatability between native
173 and invasive ranges at phenotypic and genetic levels[10,14]. We observed congruent patterns in
174 our data: among North American and European XtX-EAA outlier windows, 173 showed parallel
175 associations with the same environmental variable between ranges (significantly more than
176 would be expected by chance; hypergeometric $p = 4.39\times10^{-63}$; fig. 1C), with 19% of climate
177 adaptation candidates in Europe also candidates in North America. To account for the possibility
178 that the number of parallel windows is inflated by extended linkage disequilibrium between
179 windows (and hence represents a smaller number of loci), we combined consecutive windows,
180 and windows in haploblock regions, into single windows and repeated the analysis, in which the
181 parallelism remained highly significant (hypergeometric $p = 1.08\times10^{-63}$). Consequently, many of
182 the same regions of the genome are involved in climate adaptation in both ranges.
183
184 In addition to *ELF3*, parallel XtX-EAA outlier windows included four windows within two
185 previously identified flowering-onset QTL[42]. North American, European and parallel XtX-EAA
186 outlier windows included 17, ten, and three flowering-time pathway genes, respectively,
187 however this only represented a significant enrichment (Fisher's exact test $p < 0.05$) in North
188 America (table S7). Gene ontology terms enriched in parallel XtX-EAA windows included "iron
189 ion binding" and "heme binding" (terms relating to cytochrome P450 genes), as well as
190 "gibberellin biosynthetic process" (table S8). Some cytochrome P450 genes are involved in
191 detoxification of xenobiotic compounds and the synthesis of defense compounds, while others

192 play key developmental roles[43], including contributing to the biosynthesis of gibberellin, a
193 hormone that regulates a range of developmental events, including flowering[44].

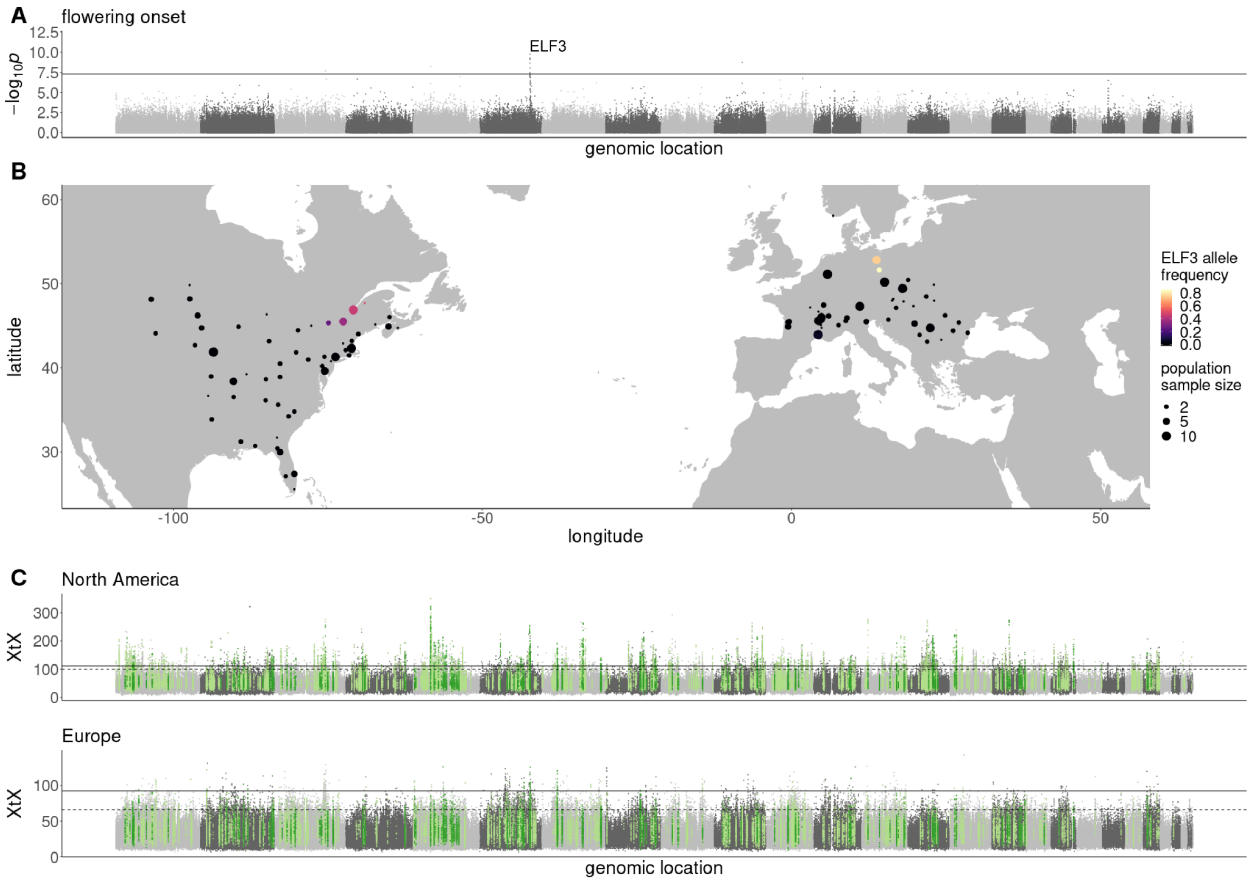

Figure 1. Signatures of climate adaptation in *Ambrosia artemisiifolia.* A. GWAS results (-$\log_{10}p$ against genomic location) for flowering onset (solid line indicates a Bonferroni-corrected p-value of 0.05). **B.** Distribution of a strongly-associated nonsynonymous SNP in ELF3 among modern *A. artemisiifolia* populations used in this study. **C.** Genome-wide XtX scans between sampling locations within each range separately. Solid lines indicate Bonferroni-corrected significance derived from XtX *p*-values; dashed lines indicate the top 1% of genome-wide XtX values. Green highlights represent the top 5% of 10kb WZA windows for each scan that are also among the top 5% of EAA WZA windows for at least one environmental variable, with dark green indicating windows shared between North America and Europe.

203 **Temporal phenotypic analysis**

204 To further identify the features of recent adaptation in *A. artemisiifolia*, we leveraged herbarium
205 samples, collected from as early as 1830, for phenotypic and genomic analyses. A trait-based
206 analysis of 985 digitized herbarium images (fig. S4) identified a significant shift in the probability
207 of flowering and fruiting over time in Europe, but this change depended on the latitude or the
208 day of the year the sample was collected (fig. S5; table S9). For the trait presence of a mature
209 male inflorescence, we identified a significant interaction between collection year and latitude
210 ($F_{1,886}$ = 7.89, p < 0.01) where in northern populations, more recently collected plants were more
211 likely to be flowering than older historic specimens. For this trait, collection day also significantly
212 interacted with collection year and more recently collected plants were more likely to be
213 flowering later in the year and less likely to be flowering earlier in the year. Similar patterns were
214 identified with the presence of fruit, as older samples were less likely to produce fruit later in the
215 season compared to recent samples (day-by-year interaction $F_{1,886}$ = 32.33, p < 0.001; fig. S5;
216 table S10). This substantial spatio-temporal change in phenology is consistent with
217 experimental common gardens that show that earlier flowering has evolved in northern
218 populations and later flowering in southern populations following the invasion of Europe[14].
219 Further, this shift in both flowering and fruit set over time supports the hypothesis that an initial
220 mismatch between the local environment and the genotypes present impacted the reproductive
221 output of *A. artemisiifolia* during the early stages of colonization, particularly in northern Europe.
222

223 **Temporal genomic signatures of selective sweeps**

224 Genome resequencing of *A. artemisiifolia* herbarium samples[11] allowed comparisons between
225 historic and modern populations. We grouped historic samples based on their age and proximity
226 to a modern population sample, resulting in five North American and seven European historic-
227 modern population pairs (table S11), which were scanned for signatures of local selective
228 sweeps by identifying windows with extreme shifts in allele frequency and extreme reductions in
229 diversity over time. We found far more evidence for recent sweeps in Europe (494 unique
230 windows) than in North America (113 unique windows; fig. S6 c.f. S7; table S11), consistent with
231 the expectation that a haphazardly-introduced invader will frequently be maladapted, initially,
232 and undergo rapid adaptation to local environmental conditions following its introduction. The
233 most dramatic selective sweep signatures were observed in Berlin and two French populations
234 in a ~16Mb region on scaffold 448 (fig. 2A; fig. S7), which accounts for 312 (63%) of the
235 European sweep windows. In Berlin, sweep windows were also observed containing and
236 surrounding the flowering onset GWAS peak that includes *ELF3*. In comparisons of spatial and

temporal signatures of selection, two and nine sweep windows were also XtX-EAA outliers in North America and Europe respectively. All such windows in Europe were in the scaffold 448 or *ELF3* regions. To further investigate the temporal shift associated with *ELF3*, we focused on the nonsynonymous *ELF3* variant across all samples within a 200km radius of Berlin (15 historic and eleven modern samples). The frequency of the variant increased from 4.9% to 73.9% between historic and modern samples, an allele frequency shift greater than 10,000 putatively neutral loci sampled from the same geographic region (fig. 2C); similar shifts were not observed in North American samples within a 200km radius of Quebec City (ten historic and eleven modern), where the *ELF3* allele is at comparably high frequencies (fig. S8).

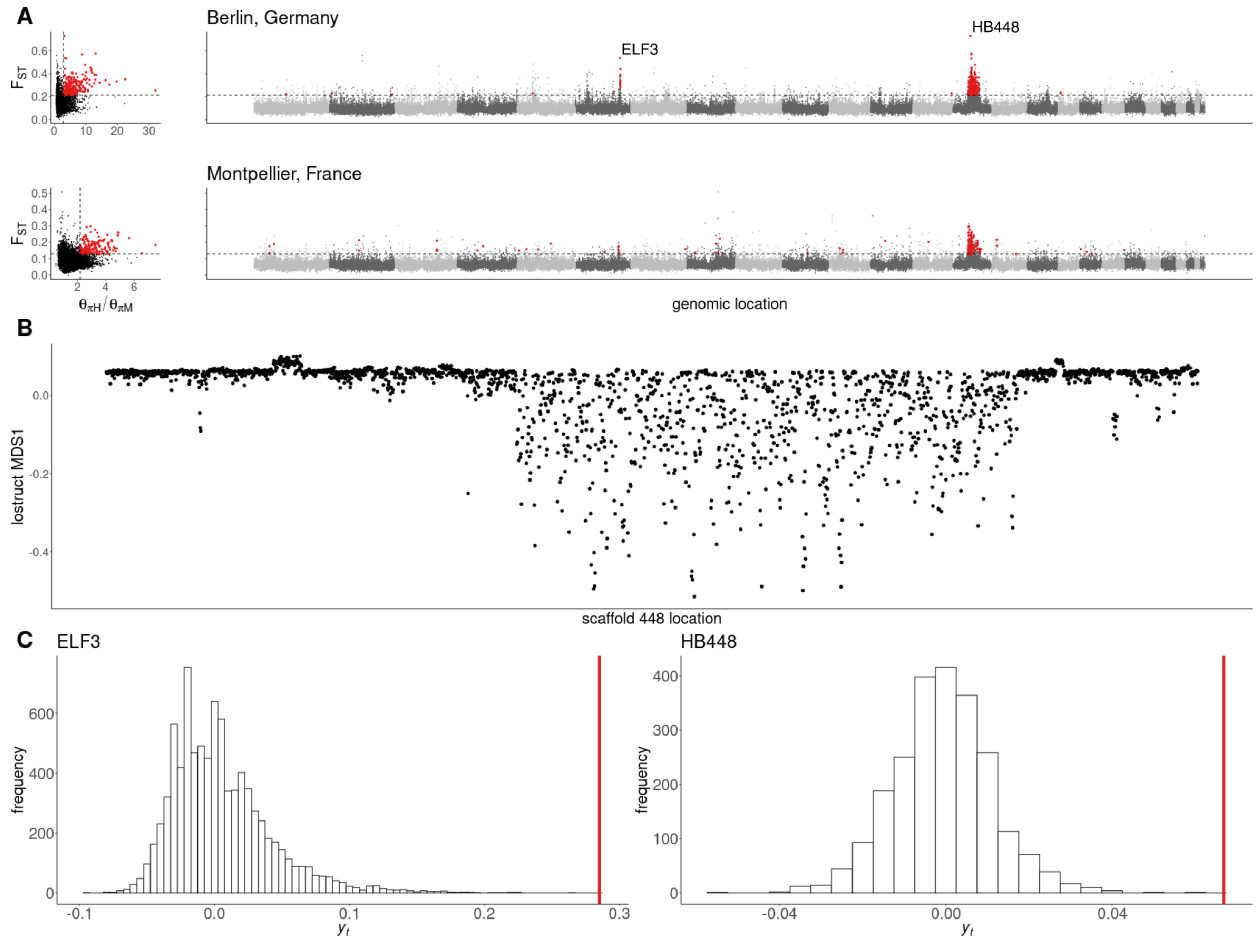

Figure 2. Temporal signatures of selective sweeps in Europe. A. Distributions of F_{ST} between historic and modern samples and the ratio of historic to modern nucleotide diversity ($\theta_{\pi H}/\theta_{\pi M}$) from Berlin and Montpellier, and F_{ST} against genomic location. Red points indicate putative selective sweep windows, which are in the top one percent of per-window F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ (dashed lines). **B.** Strong evidence for a selective sweep on scaffold 448 in European populations corresponds with local divergent population structure (MDS1), indicating the presence of a haploblock (putative chromosomal inversion) in this region. **C.** Distributions of values of a standardized measure of allele frequency change, y_t (calculated according to equation 1; histograms) for shifts between historic and modern populations across putatively neutral SNPs. y_t for scaffold 2 position 56,025,662, a nonsynonymous SNP in *ELF3* (red line), against a distribution of y_t values from 10,000 putatively neutral SNPs from samples within 200km of Berlin (left) and y_t calculated from the haplotype frequency of HB448 (red line), against a distribution of y_t values from 7,394 putatively neutral SNPs from across the European range (right).

Haploblock identification

Chromosomal inversions have previously been identified as driving local adaptation of ecotypes of *Helianthus* species[26]. We used a similar approach to identify genomic signatures of putative inversions (haploblocks) contributing to local adaptation in *A. artemisiifolia*. Briefly, we identified genomic regions in which population structure was divergent and fell into three clusters, putatively representing the heterozygous and two homozygous genotypic classes of an inversion. Further, we looked for pronounced shifts in population structure (indicating inversion breakpoints), elevated local heterozygosity in the heterozygous cluster, and increased linkage disequilibrium across the region (fig. S9). Finally, we examined mapping populations of *A. artemisiifolia*[42] for evidence of map-specific reductions in recombination across haploblock regions (i.e., suppressed recombination in haploblock regions in some maps but not others; fig. S10). This would be the pattern expected when recombination is suppressed by inversions in heterozygotes but not homozygotes, as opposed to the haploblocks being caused by global reductions in recombination in those regions. All haploblocks showed evidence of suppressed recombination in some maps but not others, with the exception of HB31 which showed suppressed recombination in all maps.

Focussing our analysis on regions showing signatures of adaptation, we identified seven haploblocks with the above genomic signatures of chromosomal inversions overlapping the 173 WZA windows that were parallel outliers for both XtX and at least one climate variable: HB2 (2.4Mb), HB5 (5.2Mb), HB21 (13.4Mb), HB27a (4.1Mb), HB27b (5.5Mb), HB31 (6.7Mb) and HB448 (16.1Mb). These haploblocks contained 31 of the 173 windows (17.9%; fig. S11), with one haploblock also corresponding to the European selective sweep region on scaffold 448 (fig. 2A;B). This suggests that these haploblock regions have played a pivotal role in generating parallel signatures of selection observed in *A. artemisiifolia*.

Haploblock frequency changes through space and time

To identify changes in haploblock frequency over time and space, which would be consistent with selection on these putative inversions, we first estimated haploblock genotypes for all historic and modern samples. Within haploblock boundaries identified using modern sample SNP data in Lostruct[45], we performed local PCAs with both historic and modern samples (table S4) and identified genotypes by kmeans clustering (fig. S9). For modern samples, we used generalized linear models to estimate the slopes of the haploblock frequencies as a function of latitude within each range. For those haploblocks that were significantly associated with latitude,

293 we compared these estimates with the genome-wide distribution of slopes for North America
294 and Europe, based on 10,000 unlinked SNPs that were randomly selected from outside
295 haploblocks and genes. The estimated slopes for haploblocks HB27a, HB27b and HB31 fell into
296 the 5% tail of the distribution for at least one of the ranges (fig. S12). However, this approach
297 did not examine temporal changes nor the combined signatures of selection over space and
298 time. To do so, we ran generalized linear models comparing haplotype frequency with latitude,
299 time (date of specimen in years) and range (North America vs. Europe; fig. 3A; fig. S13; table
300 S12-S17). All of the haploblocks showed significant changes over time, either in their average
301 frequency in one or both ranges, or in their relationship with latitude within each range, a pattern
302 that is consistent with local selection on the haploblocks. All seven haploblocks showed
303 significant associations with time or with latitude in at least one range, indicative of climate
304 adaptation. These patterns were robust to time being coded as discrete (historic vs. modern) or
305 as continuous (by year; table S12). In HB5, HB27b and HB31, we identified significant three-
306 way interactions between latitude, time and range (fig. 3A). The HB27b haplotype frequency
307 was correlated with latitude in modern and historic samples from North America, as well as in
308 modern European samples, consistent with climate-mediated selection. However, historic
309 European populations did not display an association with latitude, which may reflect
310 maladaptation during the initial stages of the European range expansion (fig. 3A; table S15,
311 S17). A significant strengthening of the relationship between latitude and haplotype frequency
312 was also observed in HB5.
313
314 To further investigate whether European populations showed evidence of recent local selection
315 on the haploblocks, we tested whether estimates of selection inferred from contemporary spatial
316 data were associated with temporal changes in haploblock frequencies between historical and
317 contemporary European populations. We used spatial variation in contemporary haploblock
318 frequencies to estimate the relative strength of local selection on these haploblocks (see
319 supplementary text S2). We specifically compared estimates of the maximum slope of latitudinal
320 clines for each putative inversion's frequency to simple population-genetic models for clines at
321 equilibrium between local selection and gene flow. In these models, cline slopes are
322 proportional to $\sqrt{s}/\sigma$, where s represents the strength of local selection for a given inversion and
323 σ is the average dispersal distance of individuals in the range[46] (supplementary text S2). While
324 our estimates of selection are, therefore, scaled by the dispersal rate, dispersal should equally
325 affect all inversions within a given range, allowing us to infer the relative strength of spatially
326 varying selection for each putative inversion. We found that estimates of the relative strengths of

local selection across the haploblocks were correlated between the ranges, indicating parallel patterns of local selection along the latitudinal gradient (r = 0.80, p = 0.03). We also found that changes in the haploblock cline slopes between historic and modern time points within Europe were significantly correlated with our estimates of the relative strength of spatially varying selection of the haploblocks across the European range (r = 0.86, p = 0.01 fig. S14). Such a pattern is consistent with a scenario in which historical European populations were not initially locally adapted (haploblock frequencies were initially far from local optima) and where the haploblocks subject to relatively strong local selection exhibited the greatest temporal changes in local frequency over the ensuing century. The same pattern was not observed in the North American native range, whose historic populations are likely to have been consistently closer to the local optima across the timescale of our analysis (r = 0.61, p = 0.15). Cline slopes for latitude were also shallower in historic relative to modern European populations ($t_{6.58}$ = -2.39, p = 0.05; mean absolute slope: historic EU = 0.05; modern EU = 0.15), but not so in North America ($t_{11.9}$ = 0.24, p = 0.81), consistent with initial maladaptation in Europe, followed by adaptation to local climates.

The HB448 haplotype frequency dramatically increased in the European invasive populations over time, consistent with strong range-wide selection in Europe (table S12). The haploblock also showed a significant range effect, with a significantly higher frequency in modern Europe compared to North America (historic and modern) and historic Europe (table S14). Latitudinal clines in frequency were also observed although the clinal patterns were similar over time and between ranges. Signatures of selective sweeps were observed in the HB448 region in multiple European populations (fig. 2A; fig. S10). To further distinguish the effects of drift from those of selection, we compared an empirical null distribution of allele frequency changes over time in Europe using 2,296 SNPs matched for allele frequency (fig. 2C). Only a single locus in this null distribution showed an allele frequency change greater than HB448. We then estimated the strength of selection for the putative inversion under scenarios of positive and balancing selection (see supplementary text S1). With positive selection for the putative inversion, the observed changes in frequency are consistent with a 2.2% difference in fitness (95% CI = {1.6%, 2.8%}) between individuals homozygous for the inversion relative to homozygotes for the alternative haplotype. Scenarios of balancing selection require stronger selection to explain the inversion frequency shift over time (see supplementary text S1). This estimate is smaller than many empirical estimates of selection on individual loci in natural populations[47]. The greater

360 timespan of our samples facilitated detection of strong signals of selection for loci that would
361 otherwise be missed by the short time periods afforded by most temporal studies.
362
363 **Biological functions of haploblocks**
364 Several analyses provide evidence for the biological function of these putatively adaptive
365 haploblocks. From genome annotations, we found that haploblocks were collectively enriched
366 for flowering-time pathway genes (Fisher's exact test $p = 0.034$), although individual
367 haploblocks were not significantly enriched (table S18). HB448 was enriched for the "recognition
368 of pollen" gene ontology term, with 28 (17%) of the genome's 167 genes annotated with this
369 term falling within this haploblock (table S19). HB27a was enriched for genes with the "pectate
370 lyase activity" term, including the top BLAST hit for *Amba1* (99.7% identity, *E*-value = 0), which
371 encodes the *A. artemisiifolia* protein responsible for the majority of allergic reactions[48]. A
372 detailed analysis of the HB27a region reveals a cluster of six closely-related pectate lyase
373 genes, which correspond with elevated XtX and XtX-EAA outlier windows in both ranges (fig.
374 S15). HB27b corresponds to a large-effect flowering time and height QTL[42]. HB2 also overlaps
375 the flowering time GWAS candidate *ELF3* (fig. 1A;B), and the nonsynonymous variant which
376 displays strong patterns in GWAS is only observed on one of the haploblock genotype
377 backgrounds. We also identified phenotypic associations with haploblocks by encoding
378 haploblock genotypes into our GWAS pipeline. Significant associations ($p < 0.05$ Bonferroni-
379 corrected for multiple testing across seven haploblocks) were observed for five haploblocks;
380 four haploblocks showed associations with traits related to flowering time (fig. 3B; table S20).

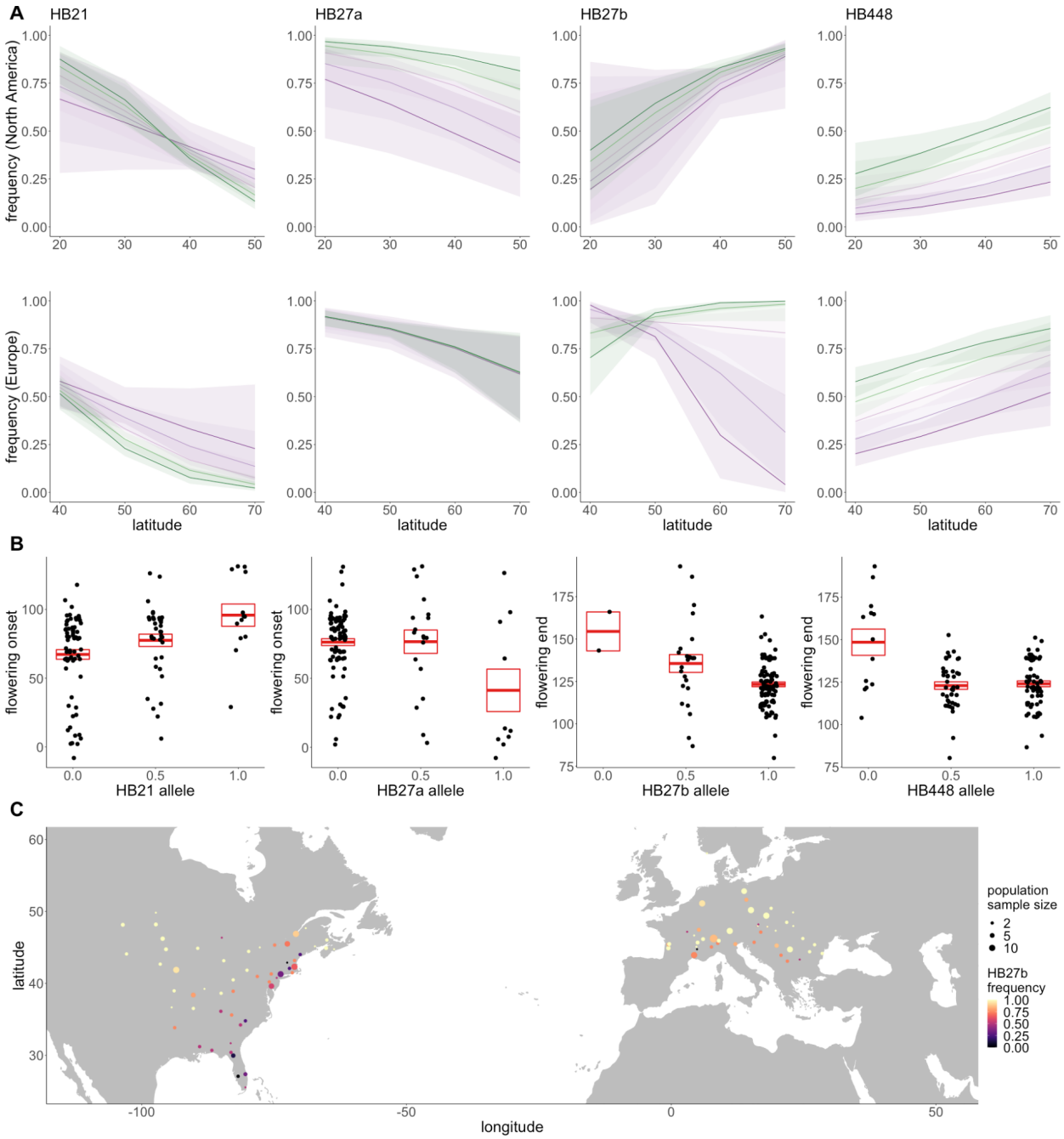

Figure 3. Haploblock distributions and trait associations. A. Logistic regression models with 95% CI ribbons (see table S12-17 for model details) of haploblock frequency (allele 1) against latitude for four haploblocks across five time bins ranging from most historic (purple) to most modern (green). **B.** Examples of significant associations between haploblock alleles and phenotypes (boxes denote mean and SEM). **C.** HB27b allele frequency in modern *A. artemisiifolia* populations.

387 **DISCUSSION**

388 We have described, at an unprecedented temporal and spatial resolution, the evolutionary-
389 genetic changes accompanying a recent and rapid invasion by a noxious pest. Our study
390 system, while unique in many ways, yields results with important general implications for our
391 understanding of the genetic basis of rapid adaptation to environmental change and the
392 pervasiveness of parallel evolution in geographically widespread species.
393
394 While invasive species are often envisaged to encounter novel selection pressures as they
395 spread across alien landscapes, they must also readapt to similar environmental variation
396 encountered in their native range, as haphazardly-introduced invaders are unlikely to be well-
397 adapted to local conditions when the invasive populations initially expand across climatic
398 gradients. Much of *Ambrosia artemisiifolia*'s European invasion lies within climatic extremes
399 encountered across its native range (fig. S3). Despite this similarity in climatic variation, the
400 patterns of parallel climate adaptation between native and invasive ranges are striking given the
401 evolutionarily recent introduction of the species into Europe. As *A. artemisiifolia*'s invasion of
402 Europe consisted of multiple introductions over a brief evolutionary time scale, these patterns
403 are likely examples of 'collateral evolution'[49], in which standing genetic variation in *A.*
404 *artemisiifolia*'s native range has been co-opted for adaptation in and across the European
405 invasive range. Parallel evolution is a hallmark of natural selection and parallel changes at the
406 genetic level point to constraints and biases in the genetic pathways to adaptation that are
407 evolutionarily achievable; when certain paths to adaptation are favored, such as when beneficial
408 variants are already present in the population as standing variants, evolution will repeatedly
409 draw on the same subset of genes to reach the same adaptive endpoints.
410
411 From herbarium specimens that were sampled throughout the course of *A. artemisiifolia*'s
412 invasion of Europe, we observed an abrupt change in flowering and fruiting over time.
413 Leveraging whole-genome sequences of herbarium samples across North America and Europe,
414 we were also able to scan populations for temporal genomic signatures of selective sweeps.
415 Although some populations have experienced shifts in ancestry over time in Europe[11], peaks
416 against the genome-wide background provide compelling evidence for rapid local adaptation in
417 European populations, with the strongest genetic signals of rapid change over time
418 corresponding to some of the strongest signatures of local adaptation in our spatial analyses,
419 particularly windows in the region of the *ELF3* gene and HB448. Further, these regions show
420 parallel signals of climate adaptation in North America and are associated with adapting traits

421 such as flowering onset. These multiple lines of evidence provide strong support that climate
422 mediated selection on phenology was pivotal in shaping the adaptive genetic landscape of *A.*
423 *artemisiifolia* in Europe.
424
425 Large haploblocks (putative inversions) contribute substantially to these genetic signals of
426 parallel adaptation. We propose that these haploblocks maintain cassettes of co-selected genes
427 that effectively segregate as single alleles of large effect[22,27], providing a genetic architecture
428 suited to local adaptation in the face of high gene flow[11,19]. Consistent with this hypothesis,
429 haploblocks are enriched for genes with particular biological functions, display associations with
430 locally-adaptive traits, and carry signals of strong selection in both the native and invasive
431 ranges. The evolution of inversions along environmental gradients has been reported in a range
432 of species[23]. However, by investigating haploblocks in an invasive plant with extensive
433 timestamped collections, we have demonstrated dramatic and adaptive evolutionary change of
434 putative inversions under natural conditions, providing compelling evidence of strong and recent
435 natural selection. These data have also allowed us to estimate selection for these variants, and
436 we have shown that haploblocks with the strongest estimates of clinal selection are driven more
437 rapidly towards their putative equilibria within the invasive range.
438
439 An important question during this era of environmental upheaval is the role of adaptation during
440 range expansion and its necessity during colonization. Through our analysis of historic samples,
441 we have shown that *A. artemisiifolia* was present in regions throughout Europe well before
442 many of these adaptive variants became locally common, suggesting the species' extensive
443 phenotypic plasticity may have facilitated its initial expansion. Strong local selection further
444 improved the match between genotypes and local environments, even appearing to affect
445 reproductive output in herbarium specimens. Many of the selected variants we identified are
446 linked to traits that are key factors in the timing, length and severity of the local pollen season
447 (e.g. days to flowering onset, days to the end of pollen production, and biomass). Consequently,
448 local adaptation has played a central role in shaping the allergy season in Europe and will likely
449 continue to be critical as climate change and continued range expansion further amplify the
450 damaging effects of this hazardous weed[50].

METHODS

Genome assembly

Seeds collected from a wild *Ambrosia artemisiifolia* population in Novi Sad, Serbia (lat. 45.25472, lon. 19.91231) were sown in potting soil at a greenhouse facility at the Ringve Botanical Garden, NTNU University Museum (Trondheim. Norway). After 160 days of growth under stable light and watering conditions, young leaf tissue from mature individual plant "NSS02/B" was sampled and flash-frozen in liquid nitrogen. These tissues were then shipped to Dovetail Genomics for high molecular weight DNA extraction and library building.

DNA samples were quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library (~20kb) for PacBio Sequel was constructed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) using the manufacturer recommended protocol. The library was bound to polymerase using the Sequel II Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel II). Sequencing was performed on PacBio Sequel II 8M SMRT cells generating 65.9 Gb of data. These PacBio CCS reads were used as an input to Hifiasm[51] with default parameters.

For each Dovetail Omni-C library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DNAse I, chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The library was sequenced on an Illumina HiSeqX platform to produce ~30x sequence coverage. Then HiRise used (See read-pair above) MQ>50 reads for scaffolding.

The HiFi assembly yielded a 2.1-Gbp genome, substantially larger than previous genome size estimates of 1.13-1.16Gbp[31,32]. We used GenomeScope 2.0 to estimate the genome size and ploidy using 21mers identified in the reads with Jellyfish 2.3.0[52]. Genomescope estimated the haploid genome size to be 1.04Gb using a diploid model (fig. S16), a better model fit (95%) than the tetraploid model (91%), which also vastly underestimated the haploid genome size (497 Mb). This finding was consistent with the smudgeplot produced by Genomescope, which also indicated diploidy (fig. S17). Genomescope also identified that the sample was highly

heterozygous (3.56%) which likely falsely-inflated the haploid genome assembly size through the inclusion of large heterozygous contigs (haplotigs) as distinct contigs. In order to remove one haplotig from each heterozygous pair, we used Purge_dups[28]. Purge_dups uses both read depth and sequence homology to identify and remove duplicated contigs or duplicated regions of a contig, and reduced our assembly size to 1.2Gbp (table S1). BUSCO version 5.1.3 (Benchmarking Universal Single-Copy Orthologs)[33] comparisons of original and reduced assemblies using the eukaryota odb10 dataset (table S1) demonstrated a two-fold reduction of duplicated BUSCO genes (from 233 to 115) with an increase in missing BUSCO genes of only zero to two.

The HiFi *de novo* assembly, purged of haplotigs, and Dovetail OmniC library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies[29]. Dovetail OmniC library sequences were aligned to the draft input assembly using bwa[53]. The separations of Dovetail OmniC read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold (fig. S18). The NCBI[54] genome submission portal identified 16 contigs containing bacterial contamination (totalling 7.3 Mb) which were subsequently removed from the final assembly.

To assess the presence of remnant haplotigs and other assembly artifacts, we mapped Illumina reads used in the reference genome assembly to the haplotig-purged reference genome FASTA file using AdapterRemoval[55], BWA-MEM[53] and Picard MarkDuplicates (https://broadinstitute.github.io/picard/), and measured average sequencing depth and heterozygosity of the alignment in non-overlapping 1Mb windows across the genome. Window depth was never greater than two times higher or 0.5 times lower than the mean, and furthermore regions of both low depth and low heterozygosity were distributed throughout the genome. The fact that there were no large regions with both low read-depth and low heterozygosity points to the success of the haplotig removal (fig. S19). Minimap2 was used to align the genome against itself, after filtering for alignments less than 20kb and fewer than 1000 matches, to identify homologous blocks that may represent haplotigs. This analysis revealed no large syntenic blocks among scaffolds, further confirming the absence of recent allopolyploidy and the success of the haplotig removal (fig. S20).

519 **Whole-genome resequencing samples**

520 Whole-genome resequencing data used in this study have previously been described in Bieker
521 *et al.*[11]. Modern samples were field-collected between 2007 and 2019, and historic samples
522 were sequenced from herbarium specimens collected between 1830 and 1973. 121 modern
523 samples with corresponding phenotype data collected by van Boheemen, Atwater and
524 Hodgins[14] were used for genome-wide association studies. 284 modern samples (from
525 populations with a sample size >= 2) were used for environmental-allele associations. 97
526 historic and 100 modern samples divided into twelve populations were used for historic-modern
527 population comparisons (table S11). For ELF3 analysis, 26 samples from within 200km of Berlin
528 (15 historic and eleven modern) and 21 samples within 200km of Quebec City (ten historic and
529 eleven modern) were used. Genotyping and analysis of haploblocks was performed using 311
530 modern and 305 historic samples. For details of each sample see table S4.

531

532 **Sample alignment, variant calling and filtering**

533 FASTQ files from historic and modern *A. artemisiifolia* samples from North America and
534 Europe[11] were aligned to our new reference genome using the Paleomix pipeline[56], which
535 incorporates AdapterRemoval[55], BWA-MEM[53], Picard MarkDuplicates
536 (https://broadinstitute.github.io/picard/) and GATK IndelRealigner[57]. Mean depths of alignments
537 ranged from 0.33X to 18.03X with a mean of 3.67X for historic samples, and 1.60X to 39.80X
538 with a mean of 6.26X for modern samples (table S4). Variants were called in the higher-depth
539 modern samples using GATK UnifiedGenotyper[58]. GATK VariantFiltration[57] and VcfTools[59] were
540 used to filter variant calls. SNP and indel calls were separately filtered using GATK hard-filtering
541 recommendations (SNPs: QD < 2.0, FS > 60.0, SOR > 3.0, MQ < 40.0, ReadPosRankSum < -
542 8.0, MQRankSum < -12.5; indels: QD < 2.0, FS > 200.0, SOR > 10.0, ReadPosRankSum < -
543 20.0, InbreedingCoeff < -0.8). Additionally, SNPs and indels were separately filtered for sites
544 with depth (DP) less than one standard deviation below the mean, and greater than 1.5
545 standard deviations above the mean. Individual genotypes were set to missing if their depth was
546 less than three, then variants with greater than 20% missing across all samples were removed.
547 Samples with greater than 50% missing variants were removed. For the remaining 311 modern
548 samples, contigs with fewer than 100 variants were removed, and the remaining 30 contigs
549 were phased and imputed using Beagle 5.2[60].

550

551 **Genome annotation**

552 To obtain RNA transcript sequences for annotation of the genome, after 160 days of growth
553 additional samples of leaf, stem, flower, root, and branch were taken from individual “NSS02/B”

554 and flash-frozen in liquid nitrogen. From these we extracted RNA from seven tissues (young
555 leaf, old leaf, stem, branch, and three stages of development of the floral head) using a
556 Spectrum Plant Total RNA Kit (Sigma, USA) with on-column DNA digestion following the
557 manufacturer's protocol. RNA extracts from all five tissues were pooled into a single sample.
558 mRNA was enriched using oligo (dT) beads, and the first strand cDNA was synthesized using
559 the Clontech SMARTer PCR cDNA Synthesis Kit, followed by first-strand synthesis with
560 SMARTScribeTM Reverse Transcriptase. After cDNA amplification, a portion of the product was
561 used directly as a non-size selected SMRTbell library. In parallel, the rest of amplification was
562 first selected using either BluePippin or SageELF, and then used to construct a size-selected
563 SMRTbell library after size fractionation. DNA damage and ends were then repaired, followed
564 by hairpin adaptor ligation. Finally, sequencing primers and polymerase were annealed to
565 SMRTbell templates, and IsoSeq isoform sequencing was performed by Novogene Europe
566 (Cambridge, UK) using a PacBio Sequel II instrument, yielding 97,819,215 HiFi reads. To
567 prepare the raw IsoSeq RNA data for downstream use in the annotation of the genome, we first
568 identified the transcripts in the PacBio single-molecule sequencing data by following the IsoSeq
569 v3 pipeline provided by PacificBiosciences (https://github.com/PacificBiosciences/IsoSeq).
570 Briefly, the pipeline takes PacBio subread files as an input and undergoes steps of consensus
571 generation, demultiplexing of primers, IsoSeq3 refinement, followed by a final clustering of the
572 reads.
573
574 Prior to annotation of the genome, repetitive elements were identified using RepeatModeler2[61].
575 ProExcluder[62] was then run to remove any protein coding genes from the repeat library.
576 RepeatMasker[63] was used to mask the genome using the finalized repeat library (table S2). A
577 large fraction of the genome consisted of repetitive sequence (67%; fig. S1). Retroelements
578 were the largest class (39.5%), with LTRs, particularly Gypsy (7.87 %) and Copia (18.98 %), the
579 most prevalent retroelements.
580
581 Genome annotation was performed using the genome annotation pipeline MAKER2 version
582 2.31.9[36] with *ab initio* and homology-based gene predictions. 3,819 unique UniProtKB/Swiss-
583 Prot protein sequences[64] from asterids (a monophyletic group of flowering plants), *Asteraceae*
584 (sunflower family), and *Ambrosia* (ragweeds) were used for homology-based gene prediction.
585 As no training gene models were available for *A. artemisiifolia*, we used CEGMA[65] to train the
586 *ab initio* gene predictor SNAP[66]. MAKER2 was run with command-line arguments
587 *model_org=simple*, *softmask=1*, *augustus_species=arabidopsis* and the *snaphmm* parameter

588 was set to the HMM generated in the manual training of SNAP. As expressed sequence tag
589 (EST) evidence, we used the IsoSeq clustered reads merged with a previously described
590 transcriptome[11].
591
592 A high confidence gene set of 29,849 gene models with strong protein or transcript support was
593 identified. These gene models had average coding lengths of 1.09 kb and 7.18 exons per gene
594 (table S3). Gene models were compared with *Arabidopsis thaliana* annotations (TAIR10
595 representative gene model proteins[67]) and the UniProtKB database using the *blastp* command
596 in BLAST+[68]. Using an E-value threshold of 1×10^{-6}, 26,688 (89.4%) genes matched TAIR10
597 annotations and 23,019 (78.1%) were matched UniProtKB (table S21). Gene ontology (GO)
598 enrichment was assessed using GO terms from *A. thaliana* TAIR 10[67] BLAST results. To
599 identify GO terms enriched among candidate lists, the R/topGO package[69] was used with
600 Fisher's exact test, the 'weight01' algorithm, and a p-value < 0.05 to assess significance.
601 Additionally, annotations were cross-referenced with 306 *A. thaliana* FLOR-ID flowering time
602 pathway genes[70]. 538 predicted *A. artemisiifolia* genes were matched to this dataset,
603 representing 212 unique FLOR-ID genes. Enrichment of flowering time genes was also
604 assessed in candidate gene lists using Fisher's exact test and a $p < 0.05$ threshold. The effects
605 of imputed variants on predicted genes were estimated using SnpEff[71].
606
607 **Allele frequency outliers and environmental allele associations**
608 Imputed genotype data from modern samples were divided for between-range and within-range
609 analyses in PLINK 1.9[72], and a minor allele frequency threshold of 0.05 was applied within data
610 subsets. For within-range analyses, sampling locations with fewer than two samples were
611 excluded and allele frequencies were calculated for each sampling location, resulting in 955,827
612 SNPs across 143 samples and 43 populations in North America and 941,649 SNPs across 141
613 samples and 31 populations for Europe. Allele frequency outliers were identified within each
614 range using the BayPass core model[39], with an Ω covariance matrix computed from 10,000
615 randomly-sampled SNPs that were located outside annotated genes and haploblocks, and
616 pruned for linkage disequilibrium using a window size of 50kb, a step size of 5bp and an r^2 of
617 0.5 in PLINK[72]. To identify allele frequency variation associated with environmental variables
618 within ranges, 19 bioclimatic variables were extracted for each sampling location from the
619 WorldClim database[40] using the R/raster package[73]. Associations with latitude and longitude of
620 sampling location, along with the 19 bioclimatic variables, were assessed for correlations with
621 population allele frequencies using Kendall's τ statistic in R[74]. Genome-wide XtX and τ results

622 were analyzed in non-overlapping 10kb windows using the weighted-Z analysis (WZA)[41], with
623 the top 5% of windows designated outliers.
624
625 **Genome-wide association studies**
626 Imputed genotypes from modern samples were filtered in PLINK 1.9[72]. Non-SNP sites and sites
627 with more than two alleles were removed. The 121 samples overlapping those phenotyped by
628 van Boheemen, Atwater and Hodgins[14] were retained (table S4), and sites with a minor allele
629 frequency below 0.05 were removed, resulting in 950,742 SNPs for analysis. Genome-wide
630 association studies (GWAS) were performed across 121 individuals from both North American
631 (n = 43) and European (n = 78) ranges using EMMAX[75], and incorporating an identity-by-state
632 kinship matrix (generated in PLINK 1.9)[72] to account for genetic structure among samples. The
633 kinship matrix was computed using 650,301 SNPs which remained after pruning for linkage
634 disequilibrium using a window size of 50kb, a step size of 5bp and an r^2 of 0.5. Candidate SNPs
635 were identified using a conservative threshold of Bonferroni-corrected p-values < 0.05.
636
637 **Phenotypic analysis of herbarium specimens**
638 We conducted a trait-based analysis of herbarium specimens found in the Global Biodiversity
639 Information Facility database (gbif.org 2021). We compiled information from all *A. artemisiifolia*
640 European herbarium specimens for which there was a digitized image of the individual in the
641 database alongside corresponding metadata (location and collection date). The collection date
642 spanned 1849 to 2020 (median 1975) and comprised 985 specimens. We determined the stage
643 of flowering (no male inflorescence present, only immature male inflorescence present, mature
644 male inflorescence present) for each image. The presence of fruit was also recorded. The male
645 inflorescence was used as an indicator of flowering as these structures are more visually
646 prominent than female flowers and the onset of male and female flowering is highly correlated[14].
647 Male florets consist of prominent spike-like racemes of male capitula, and are found at the
648 terminus of the stem, whereas female florets are observed to be in inconspicuous cyme-like
649 clusters and are arranged in groups at the axils of main and lateral stem leaves (fig. S4). The
650 dates when the specimens were collected were converted to Julian day of the year. We
651 conducted a generalized linear model with a binomial response and logit link (glm R). Both
652 binary traits (presence of a mature male inflorescence; the presence of fruit) were included as
653 response variables in two separate models. The significance of the effects were tested using the
654 Anova function (Car package R)[76] using type 3 tests. For both models, the predictors of latitude,
655 day of the year, and collection year as well as all interactions were included. Non-significant

interactions were removed in a stepwise fashion, starting with the highest order. Latitude of origin strongly correlates with flowering time in common garden experiments[14] and we expected northern populations to evolve early flowering relative to the start of the growing season to match the shorter growing seasons in these areas. As a result, if local phenology has evolved to better match the local growing season we predicted a collection year by latitude interaction, as the relationship between latitude and the probability of flowering in wild collected accessions should change over time when controlling for the day of collection.

Historic-modern genomic comparisons

To identify targets of recent selection, we compared historic and modern samples from twelve locations (five locations from North America and seven from Europe; table S11). Historic samples were grouped based on age of sample and proximity to a modern population. Analyses were performed in ANGSD[77] using genotype likelihoods. For each population location we calculated pairwise nucleotide diversity (θ_π) for historic and modern populations separately, and F_{ST} between historic and modern populations at each location. Statistics were calculated in non-overlapping 10kb windows, and windowed θ_π values were normalized by dividing by the number of sites in each window. At each location, windows with θ_π more than two standard deviations below the mean in both historic and modern populations were excluded from the analysis. We identified putative selective sweeps in each population as windows with extreme shifts over time in allele frequency as well as extreme reductions in diversity (i.e. windows in the top one percent of both F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ distributions).

Temporal allele frequency shifts in candidate loci

In order to track allele frequency shifts over time, we estimated contemporary and historical allele frequencies of the *ELF3* non-synonymous SNP and the haploblock HB448, which are two candidate loci for recent selection in Europe. Both candidates showed evidence of local selection using spatial analysis of modern populations, as well as sweep signals in temporal comparisons of individual populations. These calculations were performed in geographic regions where this recent selection is believed to have occurred at both historic and contemporary timepoints. ANGSD (*-minMapQ 10 -minQ 5 -GL 2 -doMajorMinor 1 -doMaf 2 -doIBS 1 -doCounts 1 -doGlf 2*) was used to calculate the allele frequency of the early flowering *ELF3* allele (ScBFxKa_2:56025662) in 15 historic and eleven modern samples from within 200km of Berlin, whilst the frequency of HB448 in Europe was ascertained using haploblock frequency estimates from across the European range (see below). To understand the

690 magnitude of these allele frequency shifts relative to putatively neutral alleles elsewhere in the
691 genome, we calculated a standardized measure of frequency change, y_t, using estimates of
692 historic, p_0, and contemporary, p_t, allele frequencies according to the equation:
693
694 $$y_t = \frac{p_t - p_0}{\sqrt{tp_0(1 - p_0)}}$$
695
696 where *t* is the number of generations separating the frequency estimates (equivalent to the
697 number of years due to ragweed's annual lifecycle). As we show in supplementary text S3, the
698 distribution of y_t estimates under neutrality are predictable and roughly independent of the initial
699 frequency of each neutral variant once the loci with low-frequency initial minor allele frequencies
700 are filtered out. To further assess if selection was the likely cause of temporal changes of the
701 *ELF3* and HB448 variants, we estimated the distribution of y_t estimates for computed from
702 10,000 randomly-sampled SNPs that were located outside annotated genes and haploblocks
703 and pruned for linkage disequilibrium using a window size of 50kb, a step size of 5bp and an r^2
704 of 0.5 in PLINK 1.9[72]. Prior to calculation of y_t, sampled SNPs were then filtered for a minor
705 allele frequency > 0.2 for HB448 comparisons and MAF > 0.05 for *ELF3* comparisons (due to
706 the low historic frequency of *ELF3* in historic Berlin populations). We then compared the
707 distributions to the y_t values of candidate adaptation loci to test whether candidate regions were
708 more divergent than the putatively neutral distribution. As a point of comparison we repeated
709 this analysis for the *ELF3* allele in Quebec. As in Berlin, this allele is at high frequencies, but
710 substantial temporal change was not expected as the populations were predicted to be closer to
711 the equilibrium over the temporal sampling period in the native range. Samples within 200km of
712 Quebec City (ten historic and eleven modern) were pooled at both timepoints. Allele frequency
713 changes of the 10,000 randomly-sampled SNPs and the non-synonymous *ELF3* allele were
714 assessed as above.
715
716 **Haploblock identification**
717 To identify signatures of large, segregating haploblocks across the genome, we performed local
718 windowed principal component analysis with Lostruct[45]. Using SNP data from 311 modern
719 samples, we extracted the first ten multidimensional scaling (MDS) coordinates across each of
720 the 30 imputed scaffolds (scaffolds with > 100 SNPs remaining after filtering) in windows of 100
721 SNPs. These MDS coordinates were then plotted along each scaffold to observe regions of
722 local structure, indicative of segregating haploblocks. We focused on outlier MDS signals that

723 overlapped parallel outlier windows for both XtX and at least one environmental variable, and
724 also showed well-defined boundaries indicative of chromosomal inversions. We tested for
725 additional evidence of inversions using PCA of MDS outlier regions and kmeans clustering in
726 R[74] to identify regions containing three distinct clusters representing heterozygotes and two
727 homozygotes. Additionally, we assessed heterozygosity from genotype data in each haploblock
728 region and in each modern sample, and measured linkage disequilibrium (the second highest r^2
729 value in 0.5Mb windows) across each scaffold bearing a haploblock for all modern samples and
730 for modern samples homozygous for the more common haploblock genotype using scripts from
731 Todesco *et al.*[26].
732
733 **Haploblock frequency changes over time and space**
734 For seven candidate inversions, a local PCA of each region and kmeans clustering was then
735 repeated in PCAngsd[78], so as to allow genotype estimation of these haploblocks in 305 historic
736 samples alongside the 311 modern samples. For this local PCA we used sites diagnostic of the
737 haplotypes (with F_{ST} values > 0.1 between the identified homozygote clusters in modern
738 samples, using vcftools v0.1.15 –weir-fst-pop (Auton and Marcketta, 2009) confined to the
739 haploblock coordinates) to reduce clustering ambiguity which was heightened by the low
740 coverage historic samples. We also conducted a PCA on 10,000 SNPs randomly-sampled from
741 the 311 modern genomes that were located outside annotated genes and haploblocks, and
742 pruned for linkage disequilibrium using a window size of 50 kb, a step size of 5 bp and an r^2 of
743 0.5 in PLINK[72]. Following this, we used generalized linear models (glm R) to assess how
744 haplotype frequency (binomial response) changed over time and space. A count of each
745 haplotype at a geographic location and year was the binomial response variable and time period
746 (historic or modern), range (North America or Europe), latitude, and all interactions between
747 these three main effects were used as predictors. Non-significant interactions were removed in
748 a stepwise fashion, starting with the highest order. PC1 from the PCA of 10,000 randomly-
749 sampled SNPs was included as a covariate to control for the effects of population structure on
750 haplotype frequency. We tested the significance of the effects in our model using the Anova
751 function (Car package R)[76] with type 3 tests. Significant differences among groups for means or
752 slopes were tested with the emmeans package using an FDR correction[79]. To determine if the
753 classification of samples into modern or historic timepoints influenced our results we ran a
754 second set of generalized linear models examining haplotype frequency as a function of
755 collection year, range (North America or Europe), latitude, and all interactions between these
756 three main effects as well as PC1, using the same approach as above. For interactions

757 involving two continuous variables (i.e., latitude and year) we tested if the slope estimates of
758 one variable were significant at specific values of the other using the package emmeans. This
759 allowed us to estimate when and where the haplotype frequencies were changing. The results
760 from both approaches (time as two categories or time as continuous) provided qualitatively
761 similar patterns.
762
763 We estimated the relative strength of selection on haploblocks along the latitudinal clines in
764 modern North American and European populations using slopes from logistic regressions (see
765 supplementary text S2). Specifically, we used generalized linear models to estimate the slopes
766 of the regression for each range and time point (modern or historic) combination (group). A
767 count of each haplotype at a geographic location was the binomial response variable and time
768 period (historic or modern), range (North America or Europe), latitude, and all interactions
769 between these three main effects were used as predictors. All interactions were retained in the
770 model and slopes and their confidence intervals estimated for each group using the function
771 emtrends (emmeans package R[79]; table S13). PC1 was included as a covariate to control for
772 the effects of population structure on haplotype frequency. We expected the slopes to be
773 shallower in the historic versus the modern European group, but similar across timepoints in
774 North America. To test this, we used a t-test and compared slopes for modern and historic
775 timepoints in each range. We also expected that the magnitude of change in the slope over time
776 would be the greatest in haploblocks showing the largest estimates of selection in Europe (table
777 S15). We estimated the relative strength of selection for the modern European range for each
778 haploblock and tested if the change in slope for each haploblock was correlated with this
779 estimate. We also examined if there was a correlation in the relative strength of selection for
780 modern North American and European haploblocks, which would indicate parallel selection
781 along the cline in each range.
782
783 We compared our slope estimates of the haploblocks to the genome wide distribution in each
784 range using 10,000 randomly selected SNPs outside of genes and haploblocks. We did this to
785 determine if our haploblocks showed stronger latitudinal patterns than the majority of SNPs, in
786 one or both ranges, which may be indicative of spatially varying selection. For the modern
787 samples in each range (North America or Europe), we fit a generalized linear model with latitude
788 as the only predictor. We did this for each null SNP and each haploblock that was statistically
789 associated with latitude.
790

791 **Recombination rates in haploblocks**

792 The haploblocks show multiple genomic signatures of reduced recombination. To confirm this
793 we analyzed recombination rates in genetic maps. Further if the haploblocks were caused by
794 global reductions in recombination rate (e.g., the region was found in an area with generally low
795 recombination such as a centromere), all maps should show reduced recombination rates.
796 However, if inversions were the cause, recombination would only be suppressed in genotypes
797 heterozygous for the inversion, while homozygous individuals would not show suppressed
798 recombination. To determine if there were genotype-specific reductions in recombination rate in
799 the seven candidate haploblocks, which would be consistent with inversions, we made use of
800 three previously generated genetic maps[42]. Markers were generated using genotype by
801 sequencing and alignments to the new reference. Details of the sequencing, alignments and
802 variant calling can be found in Prapas *et al.*[42]. We developed sex-specific genetic maps (i.e.,
803 maps for the maternal and paternal parent) using Lep-MAP3[80] for each scaffold of interest and
804 in each mapping population (an F1 mapping population and two F2 mapping populations).
805 Multiple maps were constructed since the haploblocks may have been segregating in different
806 frequencies in the parents of the mapping populations derived from outcrossing. Linkage map
807 construction was constrained by the physical order of the markers along each scaffold of
808 interest. Genetic distance (cM) was plotted against physical position along the chromosome for
809 each map and the intervals of the QTL and the boundaries of the haploblocks were visualized
810 and inspected for reduced recombination compared to the rest of the scaffold.

ACKNOWLEDGEMENTS

We thank Greg Owens & Michael Whitlock for discussions, and Samuel Yeaman and Sarah Otto for feedback on the manuscript. We are grateful to François Bretagnolle, Myriam Gaudeul, Heinz Mueller-Schaerer, Gerhard Karrer, and Bruno Chauvel for their assistance in obtaining many of the samples upon which this study is based, and Marie Brunier, Fátima Sánchez Barreiro, Yohann Roy, Luna Forcioli, Jacqueline Y. Lee for assistance during lab work. Some sequencing services were provided by the Norwegian Sequencing Centre, a national technology platform hosted by the University of Oslo. Some sequencing was performed by the NTNU Genomics Core Facility. Genome scaffolding was performed by Dovetail Genomics. Some analyses were performed on resources provided by Sigma2, MASSIVE M3 and ComputeCanada high performance computing platforms. We kindly thank the curators from the following herbaria for allowing us to destructively sample their precious collections: B, BR, BRNU, C, FI, G, GH, GOET, GZU, HBG, I, IASI, JE, L, LD, LY, MARS, MASS, MO, MPU, NEBC, NEU, NY, P, PH, PR, PRA, PRC, QFA, S, STU, TRH, UPS, US, W, WU.

FUNDING

This research received support from an NTNU Onsager Fellowship award, Norwegian Research Council Young Research Talents award 287327, and a SYNTHESYS Project award (www.synthesys.info, financed by European Community Research Infrastructure Action under the FP7 "Capacities" Program) to M. D. M., a FORMAS (2016-00453) & Carl Trygger Foundation for Scientific Research (grant CTS 14.425) to R. S., and an ARC DP220102362 and DP180102531 and HFSP RGP0001/2019 to K. A. H .

AUTHOR CONTRIBUTIONS

Paul Battlay: Software, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing - Review and editing, Visualization **Jonathan Wilson:** Software, Formal analysis, Investigation, Visualization **Vanessa C. Bieker:** Software, Investigation, Data curation **Chris Lee:** Investigation, Resources **Diana Prapas:** Software, Formal analysis **Bent Petersen:** Software **Sam Craig:** Investigation **Lotte van Boheemen:** Investigation, Resources **Romain Scalone:** Resources **Nissanka P. de Silva:** Software, Visualization **Amit Sharma:** Investigation, Resources **Bojan Konstantinović:** Investigation, Resources **Kristin A. Nurkowski:** Investigation, Resources **Loren Rieseberg:** Writing - Review and editing **Tim Connallon:** Methodology, Formal analysis, Investigation, Writing - Original draft, Writing -

Review and editing **Michael D. Martin:** Conceptualization, Methodology, Resources, Writing - Review and editing, Supervision, Project administration, Funding acquisition **Kathryn A. Hodgins:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing - Original draft, Writing - Review and editing, Supervision, Project administration, Funding acquisition, Visualization

COMPETING INTERESTS

The authors declare no competing interests

DATA AND CODE AVAILABILITY

Sequences used in reference genome assembly and annotation are available from NCBI under BioProject ID PRJNA819156 (to be released upon publication; reviewer link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA819156?reviewer=ppu225pmsn5tbfvf04s3tadkfj) . Reference genome FASTA and annotation GFF files are available from FigShare (to be released upon publication). Individual sample resequencing data are available from ENA under BioProject IDs PRJEB48563, PRJNA339123 and PRJEB34825.

SUPPLEMENTARY FIGURES

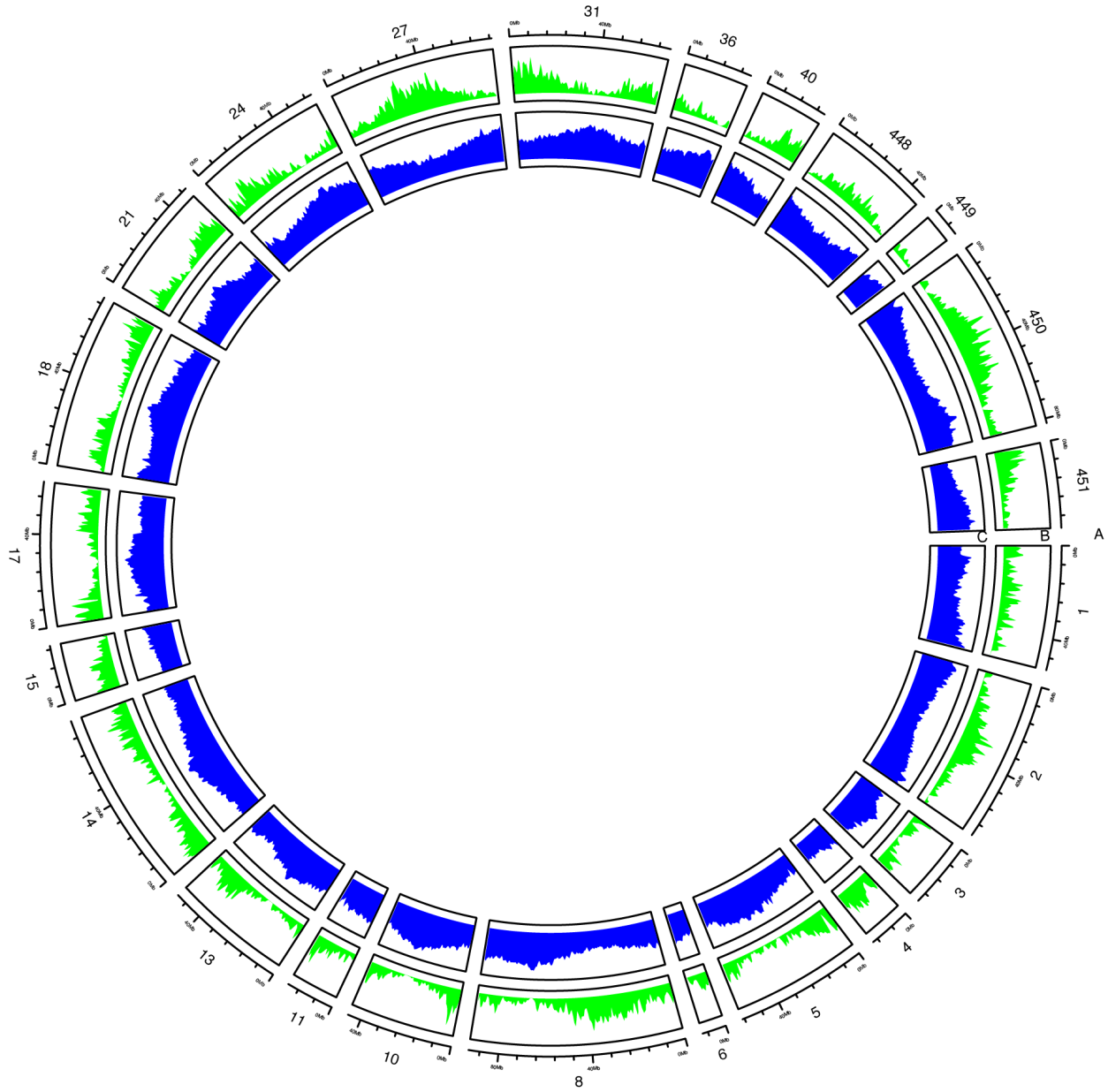

Figure S1. The *Ambrosia artemisiifolia* genomic landscape location across the 24 scaffolds greater than 10Mb in length (track **A**), and the distribution, in 1Mb windows of gene density (track **B**) and repeat density (track **C**).

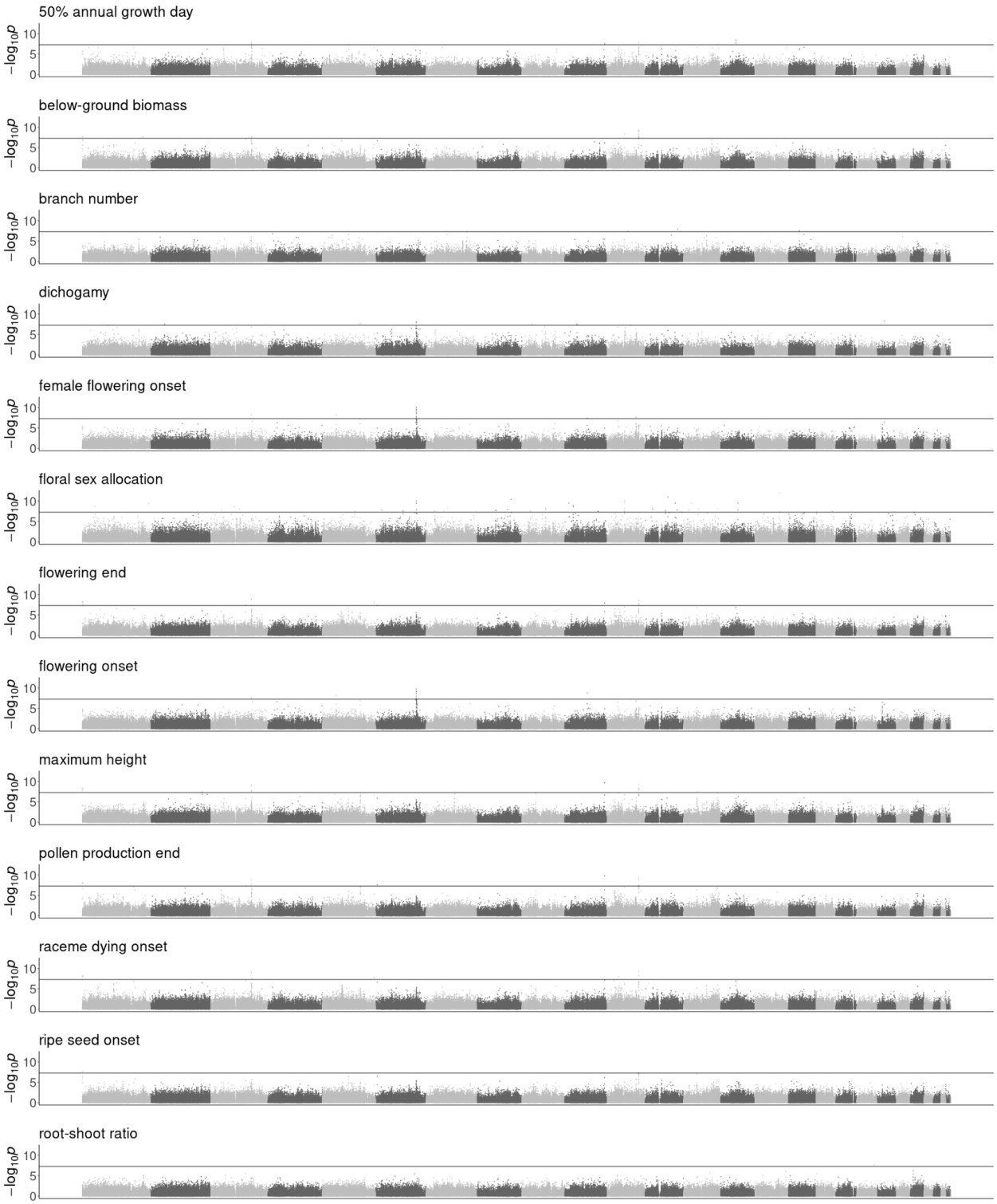

Figure S2. Genome-wide association study results for phenotypes with significant SNPs. Solid lines indicate a Bonferroni-corrected significance threshold of 0.05.

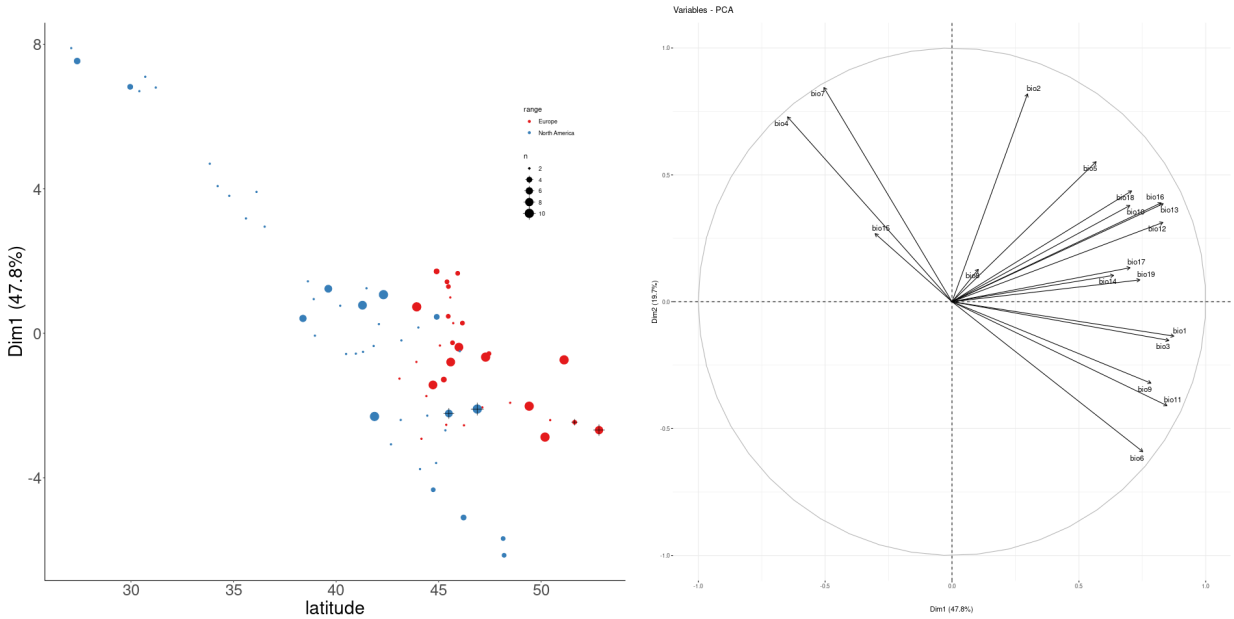

Figure S3. Environmental comparison of North American and European *A. artemisiifolia* ranges. The left pane shows the first principle component of 19 bioclimatic variables against latitude for modern ragweed populations in North America (red) and Europe (blue). Crosses identify the populations with high *ELF3* allele frequencies. The right pane shows a variable correlation plot for 19 bioclimatic variables.

Figure S4. An example herbarium specimen of *Ambrosia artemisiifolia* (left) and a detail (right) showing the mature male inflorescence (solid line) and seeds (dashed lines).

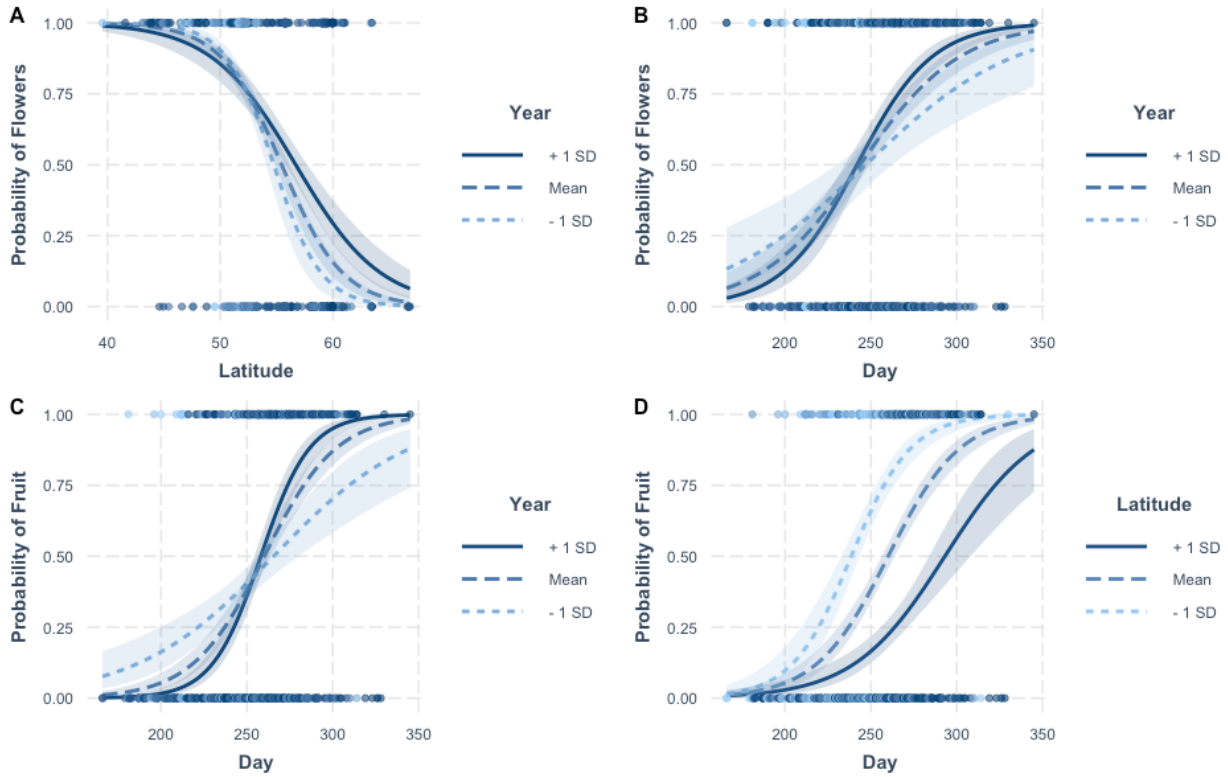

Figure S5. Interaction plots illustrating the results of generalized linear models examining the presence of mature male inflorescences (probability of flowers) or mature fruit (probability of fruit) in herbarium specimens of *A. artemisiifolia* in Europe as a function of collection day (Day), latitude of origin (Latitude) and collection year (Year). The predicted probability of observing flowers is plotted as a function of latitude (**A**), or collection day (**B**) for different collection years (mean collection year +/- 1 SD). The predicted probability of observing fruit is plotted against collection day for different collection years (mean collection year +/- 1 SD; **C**) or latitudes (mean collection latitude +/- 1 SD; **D**). Confidence intervals for the predictions are shown as are the raw data.

Figure S6. Distributions of F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ between historic and modern samples from North American populations, and F_{ST} against genomic location. Red points indicate putative selective sweep windows, which are in top one percent of per-window F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ (dashed lines).

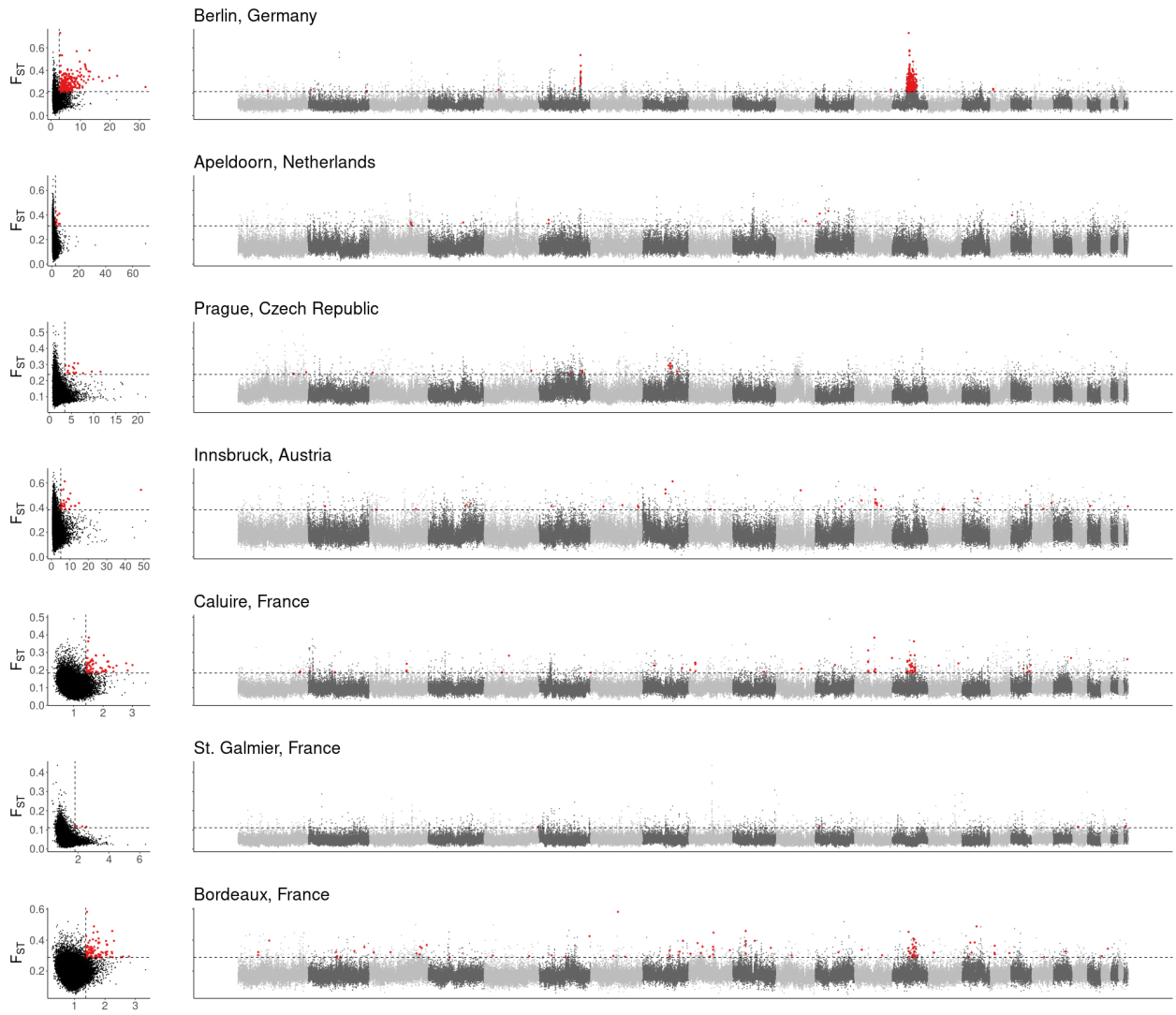

Figure S7. Distributions of F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ between historic and modern samples from European populations, and F_{ST} against genomic location. Red points indicate putative selective sweep windows, which are in top one percent of per-window F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ (dashed lines).

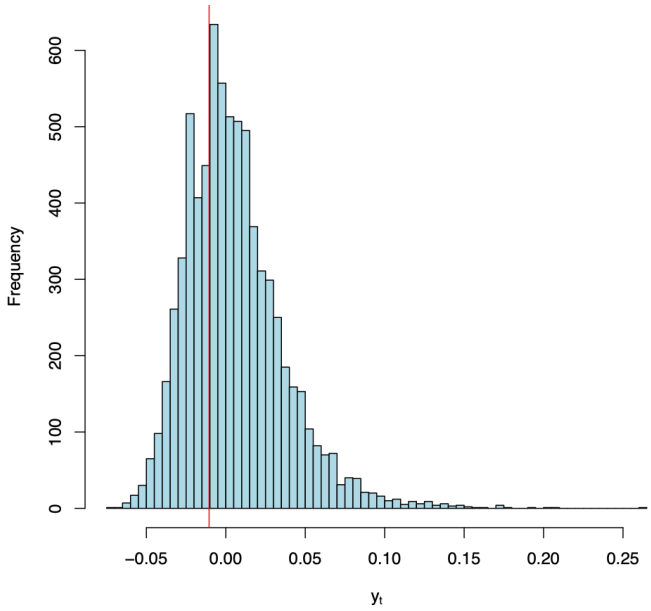

Figure S8. Distributions of y_t values (calculated according to equation 1) for allele frequency shifts in 10,000 putatively neutral SNPs (histogram) between historic and modern samples from within 200km of Quebec City. Red line indicates the y_t value for scaffold 2 position 56,025,662, a nonsynonymous SNP in *ELF3*.

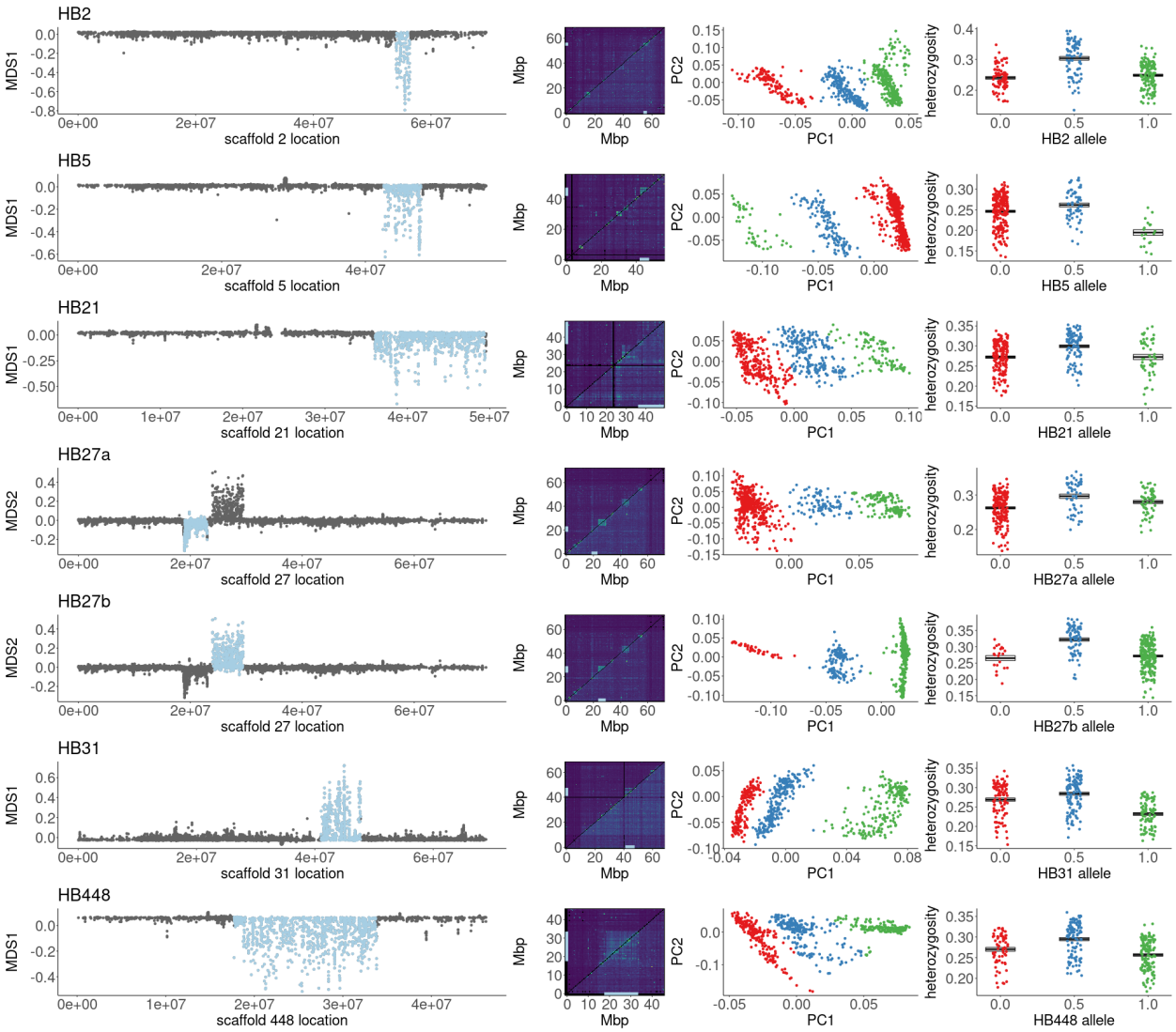

Figure S9. Haploblocks display extreme, divergent, local population structure (pale blue regions; first column). Haploblock regions (indicated by pale blue lines; second column) correspond to blocks of linkage disequilibrium (second highest r^2 in 0.5Mb windows) apparent using all modern samples (top triangle) but often reduced or absent using only samples homozygous for the more common haploblock genotype (bottom triangle). Haploblock genotypes were assigned by kmeans clustering (colours; third column) using the first two principal components of genetic variation across haploblock regions. Heterozygous haploblock genotypes show elevated mean per-site heterozygosity (fourth column; boxes denote mean and SEM).

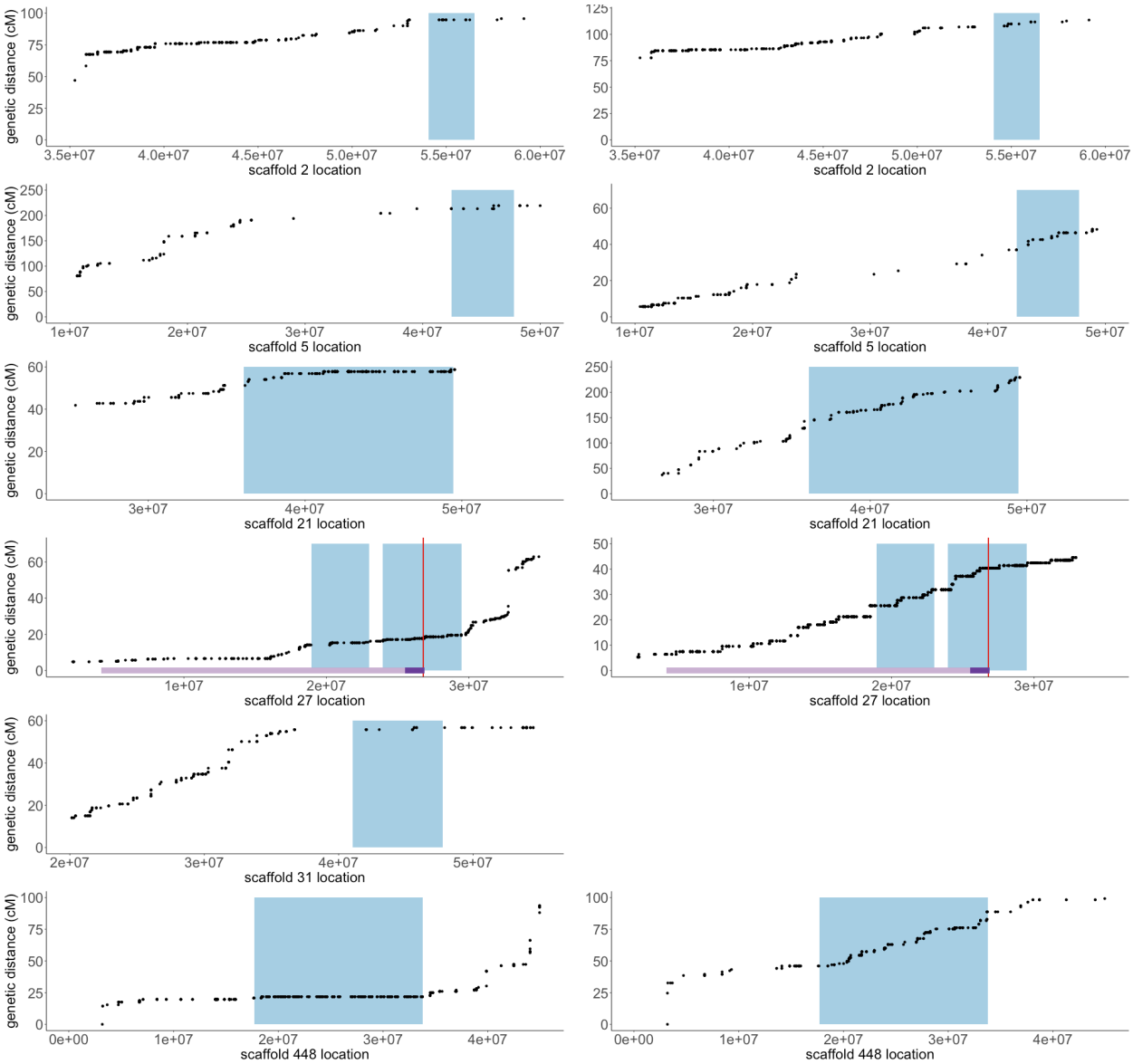

Figure S10. Evidence of genotype-specific reductions in recombination in haploblock regions. Genetic distance (cM) against physical distance (bp) along a portion of each scaffold is shown. Haploblock regions are shown in pale blue. Example maps are displayed showing both low recombination rates (left) and high recombination rates (right) for each haploblock (scaffold 31 had few markers within the haploblock for most maps, and there was no clear evidence of substantial recombination within the haploblock for any of the maps). For scaffold 27, QTL-mapping data from Prapas *et al.*[42] is shown: The marker closest to the large-effect flowering time QTL is represented by the red line. The 95% confidence intervals for male flowering time (dark purple), and height (light and dark purple) are indicated below the plots.

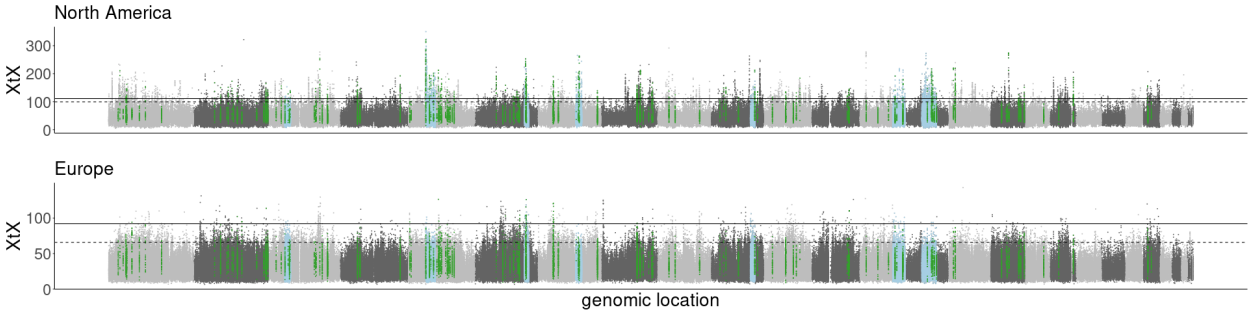

Figure S11. Genome-wide XtX scans between sampling locations within each range separately. Solid lines indicate Bonferroni-corrected significance derived from XtX *p*-values; dashed lines indicate the top 1% of genome-wide XtX values. Dark green highlights represent 10kb WZA windows with a top 5% WZA score in both ranges for XtX and at least one environmental variable. Pale blue highlights indicate haploblocks that overlap parallel WZA windows.

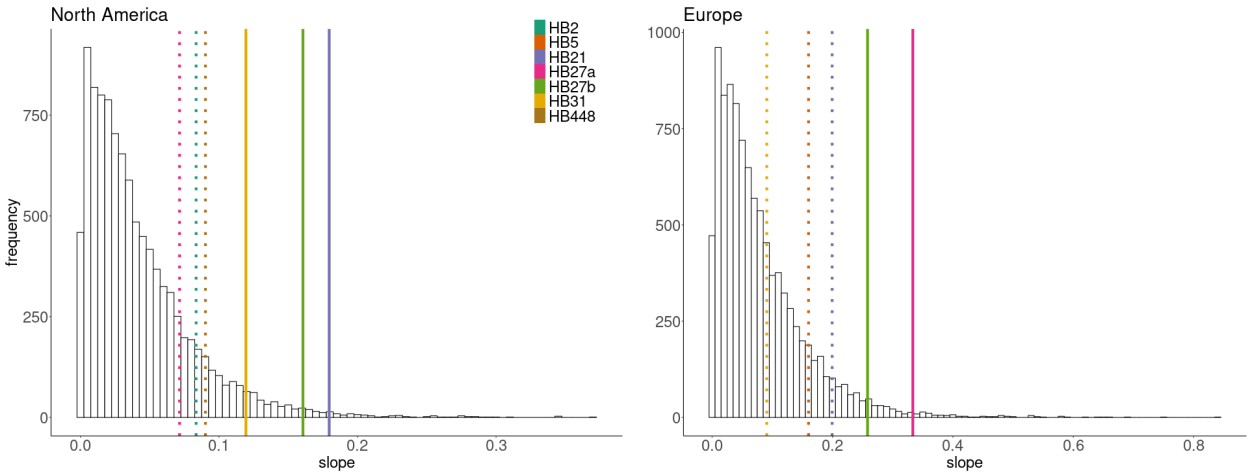

Figure S12. The distribution of slope estimates from generalized linear models of population allele counts against latitude for 10,000 randomly selected SNPs in each range. The vertical lines show the slope estimates for haploblocks with statistical associations with latitude in one range (table S12). Solid lines represent estimates in the 5% tail of each distribution while dotted lines fall below that cut-off.

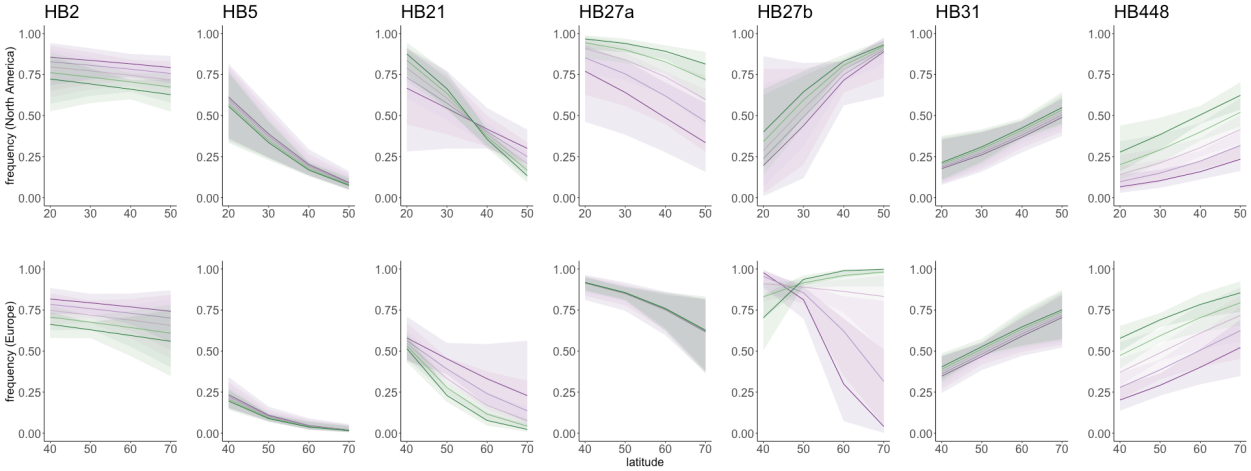

Figure S13. Logistic regression models with 95% CI ribbons (see table S12-S17 for model details) of haploblock frequency (allele 1) against latitude for each haploblock across five time bins ranging from most historic (purple) to most modern (green).

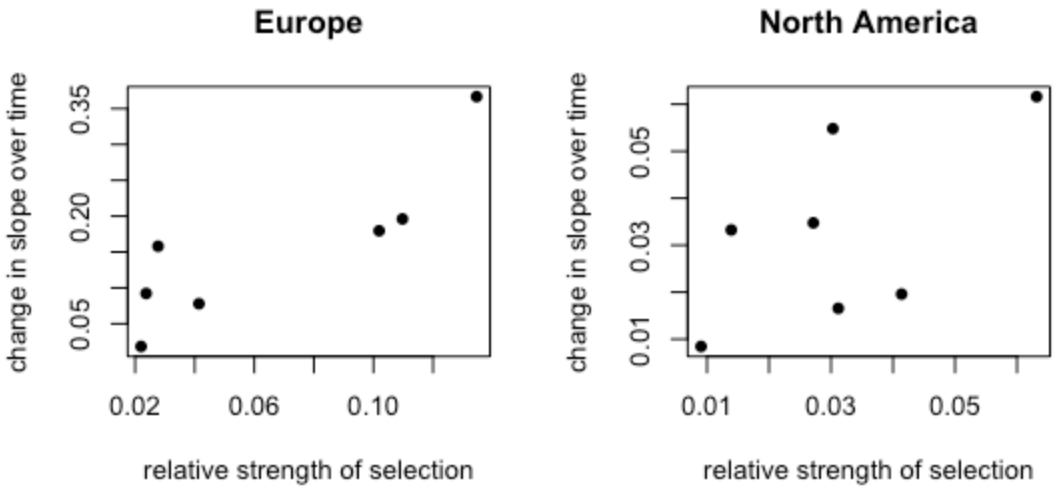

Figure S14. The change in the slope of the relationship between latitude and haplotype frequency (see table S15) between historic and modern samples compared to the estimate of selection along the latitudinal cline for the seven haploblocks (estimated from modern data in each range). A strong relationship was detected in the invasive European, but not the native North American range.

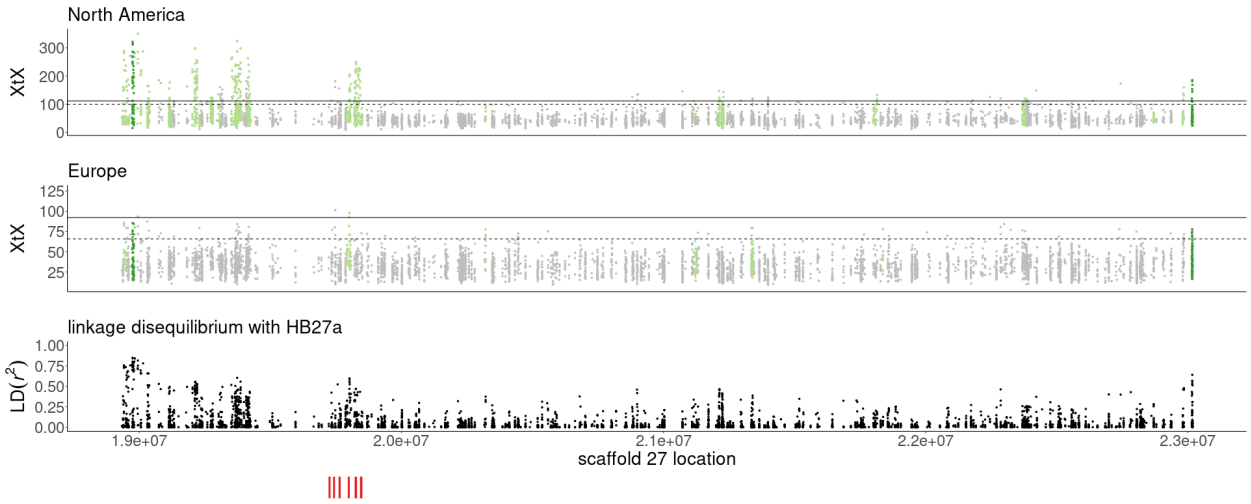

Figure S15. HB27a region (ScBFxKa_27:18930440-23019032). XtX and XtX-EAA outlier windows zoomed from fig. 1C, and linkage disequilibrium (r^2) between SNPs in the region and HB27a haploblock genotype. A cluster of pectate lyase genes, consisting of the top BLAST hit for *Amba1* and closely-related paralogues, are indicated in red below.

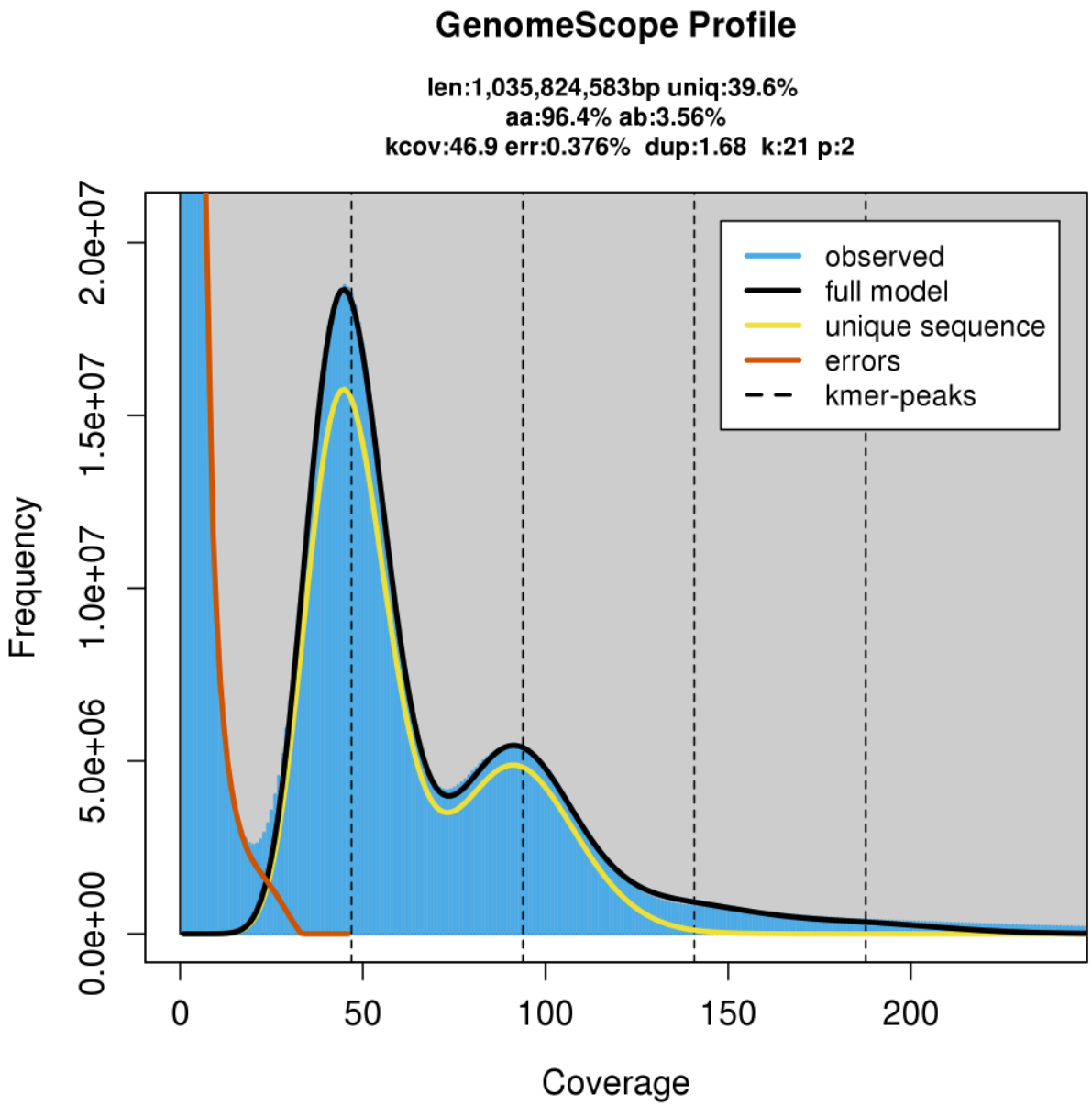

Figure S16. The kmer coverage and model fit from GenomeScope 2.0 for 21mers from the PacBio HiFi and OmniC Illumina reads for *Ambrosia artemisiifolia.*

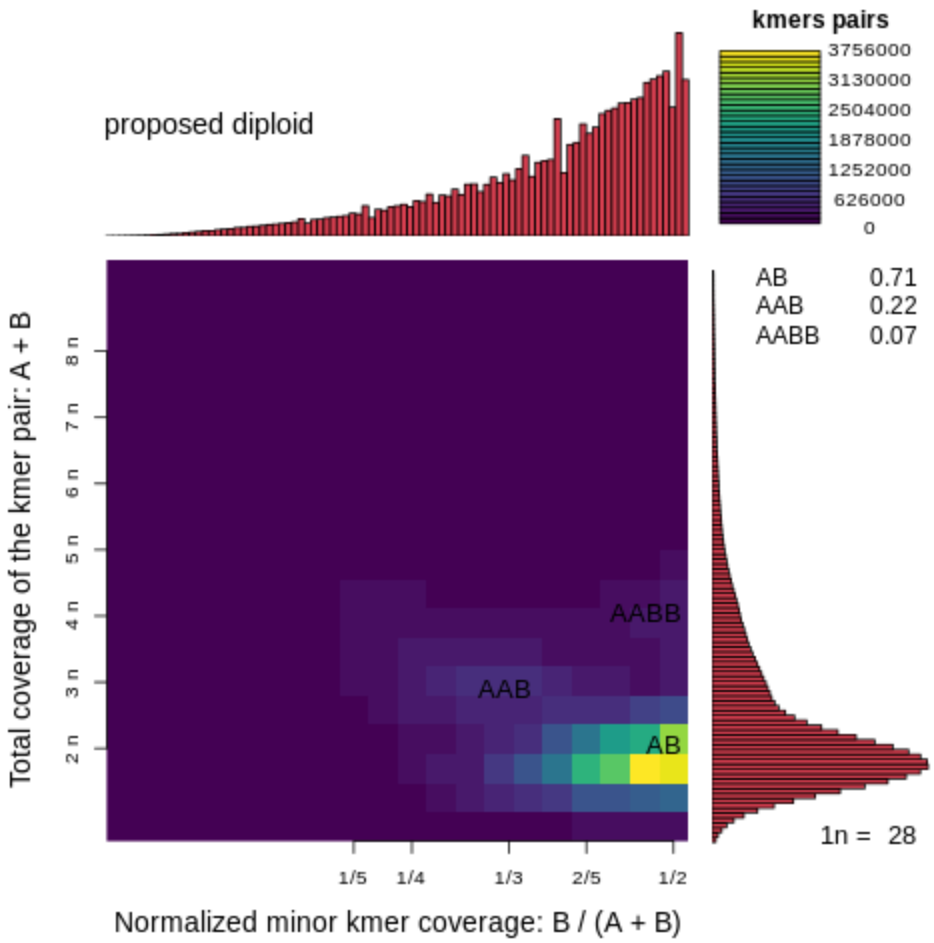

Figure S17. A smudgeplot for *Ambrosia artemisiifolia* used to estimate ploidy from heterzygous kmer pairs using 21mers from the HiFi reads.

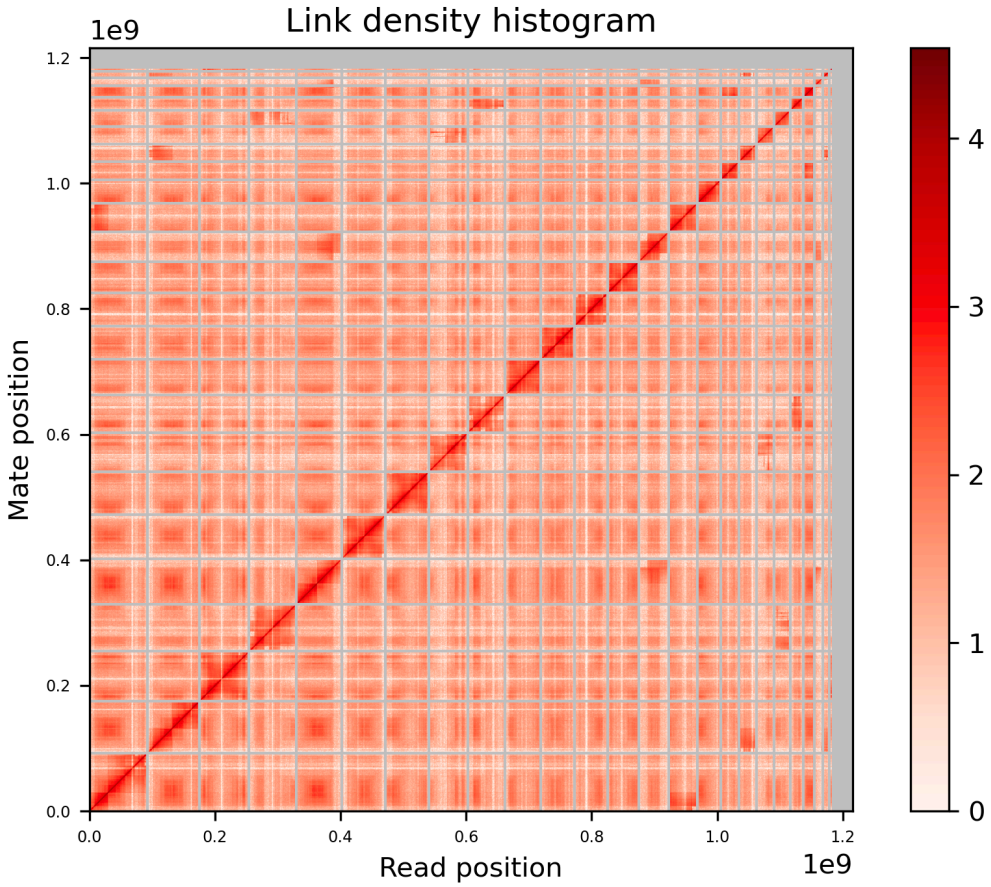

Figure S18. A link density histogram, identified by proximity ligation sequencing. The *x* and *y* axes show mapping positions of the first and second read in read pairs.

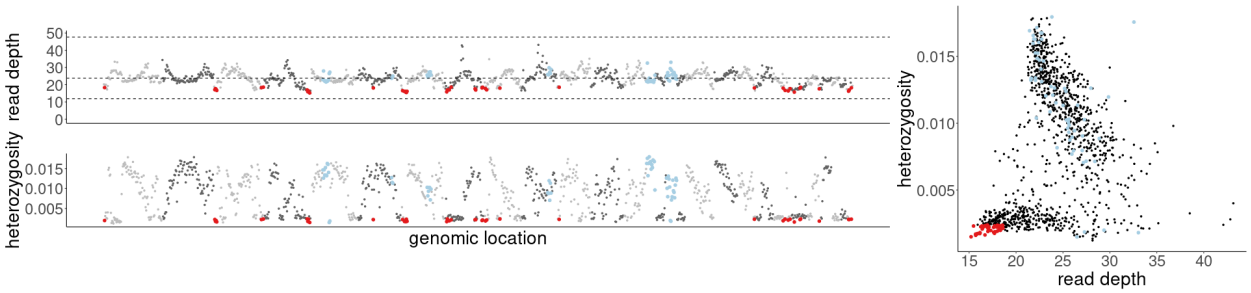

Figure S19. Read depth and heterozygosity in 1Mb windows for Illumina reads used in the reference genome assembly mapped to the final version of the reference genome. Windows in the bottom 10% of read depth and heterozygosity values are indicated in red; windows overlapping haploblocks are indicated in pale blue.

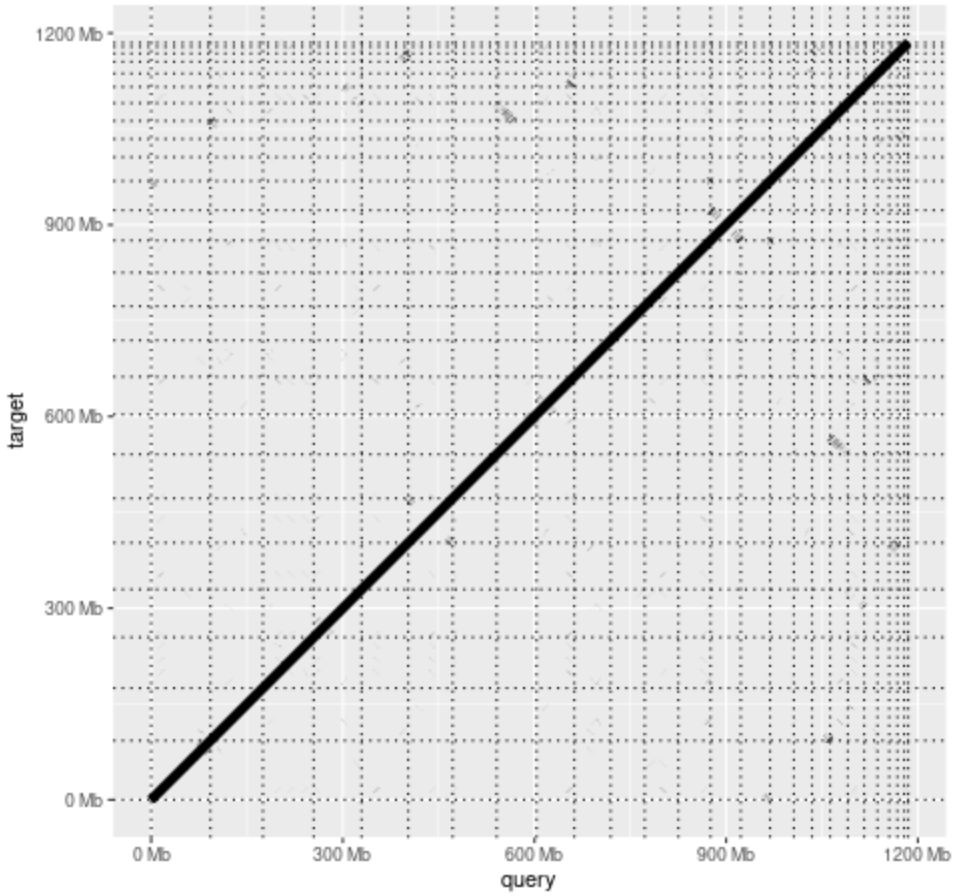

Figure S20. A dot plot of an alignment of the reference *Ambrosia artemisiifolia* genome against itself.

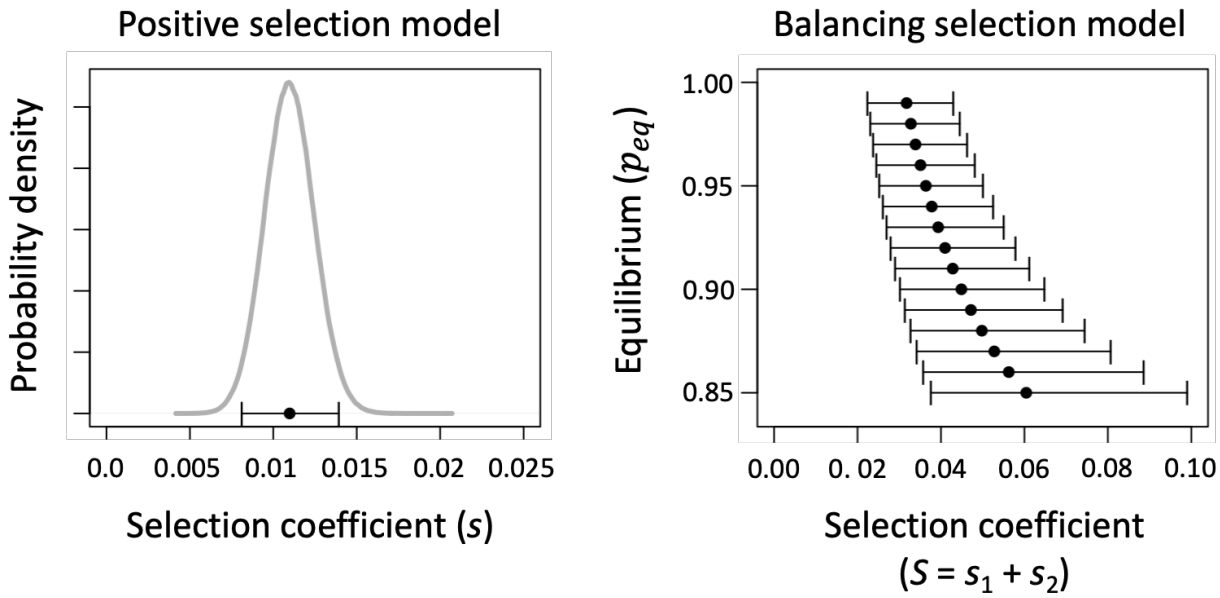

Figure S21. Selection coefficients consistent with observed frequency changes of the HB448 haplotype, under models of positive selection (left) and balancing selection (right). The left-hand panel shows the distribution of *s* values (gray) and the 95% CI for *s* (in black: parallel to the *x*-axis), consistent with temporal change in the HB448 haploblock. The distribution of s is based on simulations of 106 initial and final frequencies of the haploblock that are consistent with the estimated frequencies and error in the estimates. Eq. (S1) was use to calculate a value of s for each set of initial and final frequencies during the time interval between 1883 (early historic) and 2014 (contemporary). The right-hand panel shows the selection coefficients under the balancing selection model that are consistent with observed changes in HB448 inversion frequencies in Europe. Across the range of possible polymorphic equilibrium frequencies under balancing selection, equilibria near unity (*e.g.*, p_{eq} = 0.99) require modestly strong selection (average *S* = 0.032; 95% CI = [0.022, 0.043]) to explain the observed frequency changes in HB448, and lower equilibrium states (*e.g.*, p_{eq} < 0.90) require stronger selection.

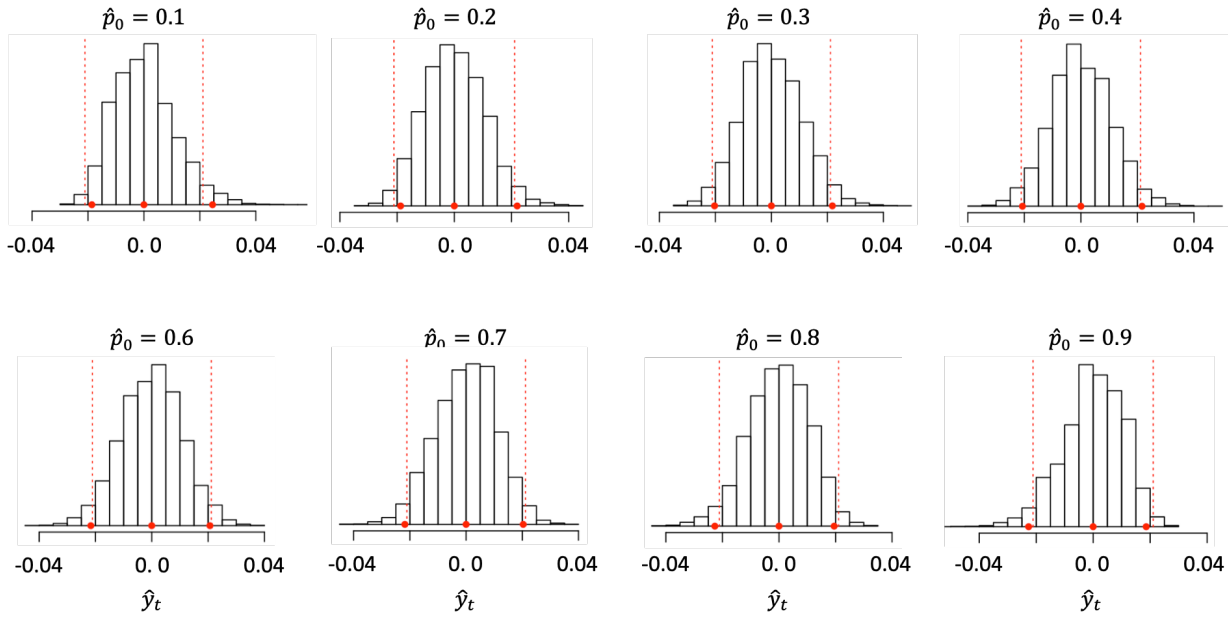

Figure S22. Simulated distribution of the scaled metric of divergence, $\hat{y}_t$, given different estimated values of the initial (historic) allele frequency. Each distribution is based on 10^4 independent and neutrally evolving SNPs. The simulations use n_0 = 182, n_t = 312, and t = 131, with a moderate effective population size (N_e = 10^4). Histograms show the distribution of $\hat{y}_t$ estimates, the red circles show the simulated mean and 95% confidence intervals for simulated data, and the vertical broken red lines show +/- 1.96 SD, where SD is the square root of our analytical expression for $var[\hat{y}_t]$.

SUPPLEMENTARY TEXT

S1: Estimating the strength of selection on HB448 using temporal changes

The putative inversion HB448 increases in frequency between historical and contemporary European populations, consistent with selection favouring its increase in the invasive range. To infer strengths of selection that are sufficient to explain the pattern of frequency change, we considered two simple, deterministic models of selection for the inversion:

- A *positive selection model* in which the inversion is favoured over the standard haplotype and is eventually expected to fix in European populations
- A *balancing selection model* in which the increased frequency of the inversion brings it closer to a hypothetical polymorphic equilibrium within the European range.

We note that, while genetic drift will inevitably play some role in the allele frequency dynamics of loci subject to selection (because populations are finite in size), evolutionary dynamics are well-approximated by deterministic models provided the allele frequencies of favourable variants are at least moderately common in the population and selection is strong relative to the inverse of the effective population size[81]. Both assumptions are easily met for the HB448 haplotype.

Positive selection model

Let p represent the frequency of the HB448 inversion haplotype. Under a model of positive selection with no dominance, the general solution for the ratio of inversion to standard haplotype frequencies (the ratio defined as $x = p/(1 - p)$) is:

$$x_t = x_0(1+s)^t$$

which is easily rearranged to solve for the frequency of the inversion:

$$p_t = \frac{p_0(1+s)^t}{1 - p_0 + p_0(1+s)^t}$$

(*e.g.* [82] pp. 200-203), where s is the fitness increase associated with each copy of the inversion (*i.e.*, the fitnesses of inversion heterozygotes and homozygotes, relative to individuals without the inversion, are $1 + s$ and $(1 + s)^2$, respectively). Under this parameterization, the difference in

relative fitness between inversion and standard haplotype homozygotes is $2s(1 + s/2) \sim 2s$, with the $2s$ approximation applying when s is small (as we infer below).

The strength of selection is a function of the inversion frequency shift from p_0 to p_t following t generations of evolution:

$$s = exp\left[\frac{1}{t} log\left(\frac{p_t(1-p_0)}{p_0(1-p_t)}\right)\right] - 1 \quad (1)$$

Since common ragweed is an annual plant, t refers to the number of years that have transpired, and p_0 and p_t can be estimated (with uncertainty) from the contemporary and historical samples.

We used eq. (1) to infer strengths of selection (s) that would be consistent with the estimated change inversion frequencies over time. We used the estimated inversion frequencies across two time points: (1) 0.4495 in the earlies historical samples with median sampling date of 1883 (n = 109 diploid individuals), and (2) 0.7724 in a contemporary sample from 2014 (n = 156). From these estimates, we simulated 10^6 inversion frequencies per time point, each based on a random sample from the binomial distribution with parameters p and $2n$ corresponding to the estimated frequency and sample size for the relevant time point (the factor of 2 in $2n$ accounts for diploidy).

We used eq. (1) to calculate 10^6 values of s that were compatible with simulated inversion frequencies for time points (1) and (2), and used these to infer the mean and 95% confidence interval for s, given our data.

Balancing selection model

Although temporal changes in the HB448 haplotype are consistent with changes predicted under the positive selection model presented above, we wished to also evaluate an alternative model in which balancing selection favours evolution of the inversion towards an equilibrium polymorphic state. To explore the strength of selection towards a hypothetical equilibrium within the European range, we considered a simple model of overdominant selection. Note that the overdominant selection model is dynamically equivalent to many other balancing selection models provided the differences in fitness among genotypes are small (consistent with our

analysis below). Our results based on the overdominance model should, therefore, apply more broadly to other scenarios of balancing selection, including scenarios involving negative frequency-dependence and antagonistic pleiotropy[83,84].

Following standard theory (*e.g.*, [85] pp. 270-272), the expected change in frequency over a generation (generation t to generation $t + 1$) is:

$$\Delta p_t = p_{t+1} - p_t = \frac{(s_1 + s_2)p_t(1 - p_t)(p_{eq} - p_t)}{1 - p_{eq}(1 - p_{eq})(s_1 + s_2) - (p_t - p_{eq})^2(s_1 + s_2)}$$

where s_1 and s_2 refer to the fitness costs of being homozygous for inversion and standard haplotypes, respectively, and $p_{eq} = s_2(s_1 + s_2)^{-1}$ is the equilibrium frequency of the inversion. Using a continuous-time approximation, we can solve for the overall selection coefficient, $S = s_1 + s_2$, that is consistent with a frequency shift from p_0 to $p_t < p_{eq}$ across t generations:

$$t = \int_{p_0}^{p_t} \frac{1 - Sp_{eq}(1 - p_{eq}) - S(x - p_{eq})^2}{Sx(1 - x)(p_{eq} - x)} dx$$

$$= \frac{\left(p_{eq} - Sp_{eq}(1 - p_{eq})\right) log\left(\frac{1 - p_t}{1 - p_0}\right) + (1 - p_{eq})(1 - Sp_{eq}) log\left(\frac{p_t}{p_0}\right) - \left(1 - Sp_{eq}(1 - p_{eq})\right) log\left(\frac{p_t - p_{eq}}{p_0 - p_{eq}}\right)}{Sp_{eq}(1 - p_{eq})}$$

Solving for S, gives us:

$$S = \frac{p_{eq}\ log\left(\frac{1 - p_t}{1 - p_0}\right) + (1 - p_{eq}) log\left(\frac{p_t}{p_0}\right) - log\left(\frac{p_{eq} - p_t}{p_{eq} - p_0}\right)}{p_{eq}(1 - p_{eq})\left(t + log\left(\frac{1 - p_t}{1 - p_0}\right) + log\left(\frac{p_t}{p_0}\right) - log\left(\frac{p_{eq} - p_t}{p_{eq} - p_0}\right)\right)} \quad (2)$$

To infer the strength of selection (S) that would be consistent with observed inversion frequencies and a given equilibrium value (p_{eq}), we simulated 10^6 inversion frequencies consistent with the estimated frequency and its sample size at historical time point (~1883) and 10^6 frequencies consistent with the estimate for the contemporary sample. (Frequencies were simulated as described in the positive selection model section, above). We used each pair of simulated inversion frequencies and eq. (2) to infer the value of S consistent with the frequency

values. The resulting distribution of 10^6 simulated *S* values was used to calculate 95% CI for *S* consistent with the data. We focused on equilibrium values outside of the 95% CI for contemporary inversion frequencies (*i.e.*, values of p_{eq} between 0.85 and 1). The results show that plausible selection coefficients under scenarios of balancing selection are consistently greater than those of the positive selection model (fig. S21). Selection under the positive selection model can, therefore, be regarded as a lower bound for the strength of selection consistent with the observed temporal changes in European HB448 inversion frequencies.

S2: Selection estimated from spatial changes in haploblock frequency

Cline theory

We will consider the simplest possible population genetics model of local adaptation in a species that is continuously distributed along a single axis of space (*e.g.*, from north to south), with x represent location along the axis, and $x = 0$ representing a point in space where the environment relevant to selection at a focal locus—in this instance, a genomic region segregating for an inversion—changes abruptly. We assume that the inversion is favoured in locations where $x > 0$ (*e.g.*, in the north) and the standard haplotype is favoured in locations where $x < 0$ (*e.g.*, the south). We further assume that population density is uniform across the spatial gradient (at least within the vicinity of the environmental transition), and that individual dispersal follows a symmetric, Gaussian distribution with variance of σ^2 (the unit of distance is arbitrary, though σ and x should have the same units, *e.g.*: if distance in x is measured in kilometres then σ should also be expressed in km; σ^2 corresponds to the migration rate, m, between adjacent patches in discrete stepping stone models)[86].

Given the stated assumptions, the inversion frequency dynamics at location x can be described using the following reaction diffusion equation:

$$\frac{dp(x)}{dt} = \frac{\sigma^2}{2}\frac{d^2p(x)}{dx^2} + \Delta p_{sel}(x)$$

where $\Delta p_{sel}(x)$ is the local response to selection[46,87]. With symmetrical strengths of selection at each side of the environmental transition, and no dominance, then $\Delta p_{sel}(x) \approx sp(x)\big(1 - p(x)\big)$ within the northern region of the range where the inversion is favoured, and $\Delta p_{sel}(x) \approx -sp(x)\big(1 - p(x)\big)$ in the southern portion of the range where the standard haplotype is favoured; both expressions are valid for modest-to-weak selection ($0 < s < \sim 0.1$). As in the positive selection model presented above, this parameterization leads to local fitness differences of ~2s between inversion and standard haplotypes. Incorporating dominance does not change our results provided the dominance relations between the alleles are consistent across the range (*i.e.*, under "parallel dominance"[88]). At equilibrium between selection and migration, the maximum cline slope will be:

$$\left(\frac{dp(x)}{dx}\right) = \sqrt{\frac{s}{3\sigma^2}} \tag{3}$$

Following Roughgarden[87], the equilibrium general solution for the cline is:

$$p(x) = -\frac{1}{2} + \frac{3}{2}\left[tanh\left(x\sqrt{\frac{s}{2\sigma^2}} + \left(\sqrt{\frac{2}{3}}\right)\right)\right]^2 \; for\; x > 0$$

$$p(x) = \frac{3}{2} - \frac{3}{2}\left[tanh\left(-x\sqrt{\frac{s}{2\sigma^2}} + \left(\sqrt{\frac{2}{3}}\right)\right)\right]^2 \; for\; x < 0$$

Estimating cline slopes by logistic regression

A logistic regression model for inversion frequency as a function of geographic location (x) is:

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The parameters of the model can be estimated by fitting the data to the log-odds (logit):

$$log\left(\frac{f(x)}{1 - f(x)}\right) = \beta_0 + \beta_1 x$$

Using the theoretical cline functions (above) to calculate the log-odds, we obtain the following slopes. For shallow clines—those with a geographically broad clinal region, where the maximum slope can be accurately estimated—we have:

$$\frac{d\; log\left(\frac{p(x)}{1 - p(x)}\right)}{dx} = 4\sqrt{\frac{s}{3\sigma^2}}$$

For steep clines—those with a narrow clinal region, where the maximum slope will be underestimated using the logit function—we have:

$$\frac{d\ log\left(\frac{p(x)}{1-p(x)}\right)}{dx} = \sqrt{\frac{2s}{\sigma^2}}$$

We get the following estimates from these two limits:

$$\frac{\sqrt{3}\beta_1}{4} \leq \frac{\sqrt{s}}{\sigma} \leq \frac{\beta_1}{\sqrt{2}}$$

Given the point estimates and 95% CI for β_1 in table S12, we can calculate plausible ranges for the lower bound of $\frac{\sqrt{s}}{\sigma}$ by multiplying the values for β_1 by $\frac{\sqrt{3}}{4}$

Comparisons of spatially varying selection among haploblocks

All of these estimates rely on the assumption that the system is at equilibrium within each range and time point, though that assumption may be more valid for some cases then others. To the extent that it is a reasonable assumption, and if dispersal (gene flow) is consistent across ranges and times, we can estimate the relative strength of selection in different contexts as:

$$\frac{s_A}{s_B} = \left(\frac{\beta_{1,A}}{\beta_{1,B}}\right)^2$$

where *A* and *B* refer to different ranges or time points.

S3: A simple null model of temporal allele frequency changes under drift

To evaluate whether temporal changes in candidate loci exceeded neutral expectations under drift in the absence of selection, we compared the distribution of the following standardized measure of divergence for a large sample of putatively neutral SNPs with the same metric calculated for selection candidates. Let divergence after *t* generations be defined as:

$$y_t = \frac{p_t - p_0}{\sqrt{tp_0(1-p_0)}}$$

where p_0 and p_t represent the initial and final frequencies of an allele at a bi-allelic locus. We shall show below that, provided loci with low minor allele frequencies are first filtered out of the analysis, the metric follows a symmetric distribution that is approximately independent of the initial frequency.

For a locus with initial frequency of p_0, the frequency after one generation of drift is given by:

$$p_1 = \frac{x}{2N_e}$$

where x is a random variable drawn from a binomial distribution with parameters $2N_e$ and p_0, where N_e is the effective population size (this follows the standard, Wright-Fisher model of genetic drift). The expected value and the variance for p_1 is therefore p_0 and $p_0(1 - p_0)/2N_e$, respectively. The model can be extrapolated for a modest number of generations, after which the allele frequency (p_t after t generations) has an expected value of p_0 and variance of $tp_0(1 - p_0)/2N_e$. The latter will eventually break down as t increases, but it should be appropriate provided $t/2N_e$ is small and the initial frequency is not too close to zero or one, as we assume below. From these expressions, the standardized measure of allele frequency divergence in the population under drift (and no selection) has an expectation of zero and a variance of:

$$var(y_t) = \frac{var(p_t)}{tp_0(1-p_0)} = \frac{1}{2N_e}$$

which is independent of the initial frequency.

In reality, error in the estimates of p_0 and p_t will also affect the test statistic, and this will tend to inflate the variance, but doesn't alter the conclusion that (under a null model of drift with no selection) the distribution of the estimates of y_t (which we denote as $\hat{y}_t$) will be roughly independent of the initial allele frequencies in the historic sample. If we define $\hat{p}_t$ and $\hat{p}_0$ is the estimates of the allele frequencies, then our test statistic is:

$$\hat{y}_t = \frac{\hat{p}_t - \hat{p}_0}{\sqrt{t\hat{p}_0(1-\hat{p}_0)}}$$

The mean and variance of $\hat{y}_t$ can be calculated using the following steps:

Step 1. The expected value and variance of $\hat{p}_t$ conditioned on the final population frequency (*i.e.*, the true frequency, p_t) is:

$$E[\hat{p}_t|p_t] = p_t$$

$$var[\hat{p}_t|p_t] = \frac{p_t(1-p_t)}{n_t}$$

where n_t represents the number of genes sampled in the contemporary population (*e.g.*, for HB448, 156 individuals were sampled for the contemporary estimate in Europe; given diploidy, we have n_t = 312).

Step 2. The expected value and variance of $\hat{p}_t$ conditioned on the initial population frequency (*i.e.*, the true frequency, p_0) is:

$$E[\hat{p}_t|p_0] = E([\hat{p}_t|p_t]\,|p_0) = E(p_t|p_0) = p_0$$

$$var[\hat{p}_t|p_0] = E(var[\hat{p}_t|p_t]\,|p_0) + var([\hat{p}_t|p_t]\,|p_0) = E\left(\frac{p_t(1-p_t)}{n_t}|p_0\right) + var(p_t|p_0)$$

$$var[\hat{p}_t|p_0] = \frac{E(p_t|p_0)\,(1-E(p_t|p_0)\,) - var(p_t|p_0)}{n_t} + var(p_t|p_0) = \frac{p_0(1-p_0)}{n_t} + \frac{tp_0(1-p_0)}{2N_e}\left(1-\frac{1}{n_t}\right)$$

where N_e is the effective population size, and n_0 is the number of genes sampled in the historic population (*e.g.*, for HB448, 91 individuals were sampled for the contemporary estimate in Europe; given diploidy, we have n_0 = 182).

Step 3. Among loci with an initial frequency estimate of $\hat{p}_0$, the true initial frequency (p_0) will, roughly, follow a distribution with mean and variance of $\hat{p}_0$ and $\hat{p}_0(1-\hat{p}_0)n_0^{-1}$, respectively. Consequently, the expected value and the variance of $\hat{p}_t$ conditioned on the initial frequency estimate $\hat{p}_0$ will be:

$$E[\hat{p}_t|\hat{p}_0] = E\{[\hat{p}_t|p_0]\,|\hat{p}_0\} = E\{p_0|\hat{p}_0\} = \hat{p}_0$$

$$var[\hat{p}_t|\hat{p}_0] = E\{var[\hat{p}_t|p_0]\,|\hat{p}_0\} + var\{E[\hat{p}_t|p_0]\,|\hat{p}_0\} = E\left\{\frac{p_0(1-p_0)}{n_t} + \frac{p_0(1-p_0)}{2N_e}t\left(1-\frac{1}{n_t}\right)|\hat{p}_0\right\} + var\{p_0\,|\hat{p}_0\}$$

$$var[\hat{p}_t|\hat{p}_0] = \hat{p}_0(1-\hat{p}_0)\left[\frac{1}{n_t} + \frac{1}{2N_e}t\left(1-\frac{1}{n_t}\right)\right]\left(1-\frac{1}{n_0}\right) + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}$$

Therefore, the expected value and the variance for $\hat{y}_t$, given an initial frequency estimate of $\hat{p}_0$, will be:

$$E[\hat{y}_t|\hat{p}_0] = E\left[\frac{\hat{p}_t - \hat{p}_0}{\sqrt{t\hat{p}_0(1-\hat{p}_0)}}|\hat{p}_0\right] = 0$$

$$var[\hat{y}_t|\hat{p}_0] = \frac{var[\hat{p}_t|\hat{p}_0]}{t\hat{p}_0(1-\hat{p}_0)} = \frac{\left[\frac{1}{n_t} + \frac{1}{2N_e}t\left(1-\frac{1}{n_t}\right)\right]\left(1-\frac{1}{n_0}\right) + \frac{1}{n_0}}{t}$$

Note that the final expressions are, once again, independent of the initial frequency, though (once again) the pathway to these results requires that $\hat{p}_0$ is not too close to zero or one. Because of this independence, we can pool loci with different initial frequency estimates (with pooling after loci with low minor allele frequency are first removed) to approximate the null distribution for $\hat{y}_t$ as well as the variance of the test statistic: $var[\hat{y}_t] = var[\hat{y}_t|\hat{p}_0]$.

Incidentally, the expression for $var[\hat{y}_t|\hat{p}_0]$ can be rearranged by solving for the effective population size across the *t* generations, *i.e.*:

$$N_e = \frac{\frac{1}{2}t\left(1-\frac{1}{n_t}\right)\left(1-\frac{1}{n_0}\right)}{tvar[\hat{y}_t] - \frac{1}{n_0} - \frac{1}{n_t}\left(1-\frac{1}{n_0}\right)}$$

A rough estimate of N_e can be obtained from a set of independent neutral SNPs by using the above formula with the estimated variance of $\hat{y}_t$ substituted for $var[\hat{y}_t]$.

Simulations

We carried simulations to test the theoretical predictions of the neutral model presented above, and find that they work well as long as initial allele frequency estimates are not too close to zero or one ($0.1 < \hat{p}_0 < 0.9$ performs well and $0.2 < \hat{p}_0 < 0.8$ is excellent). Simulations for a given value of $\hat{p}_0$ were carried out using the following steps. First, we used rejection sampling to simulate a distribution of initial population frequencies (p_0) for a given value of $\hat{p}_0$. For each SNP, we sampled a true population frequency (p_0) from a neutral stationary distribution (*i.e.*, a single draw from a symmetric beta distribution with parameters theta = 0.05, which corresponds to the population-scaled mutation rate for the locus)[89]. We then generated a frequency estimate for the SNP from a single draw from a binomial distribution with parameters p_0 and n_0 = 182, where n_0 is the number of genes sampled in the historic population. We retained the first 10^4 simulated SNPs whose estimate after binomial sampling matched the focal value of $\hat{p}_0$. From the retained SNPs, we carried out forward Wright-Fisher simulations under pure drift for *t* generations to determine the contemporary population frequency (p_t) for each SNP. We then carried out a second round of binomial sampling (with parameters p_t and n_t) for each SNP to generate a final allele frequency estimate. The frequency estimates were used to calculate $\hat{y}_t$ for each simulated SNP. Fig. S22 shows simulated distributions of $\hat{y}_t$ for different values initial frequency estimates ($\hat{p}_0$). The distributions are roughly independent of $\hat{p}_0$ and their 95% CI are well-approximated by the 95% confidence interval predicted by a normal distribution with variance corresponding to our analytical expression for $var[\hat{y}_t]$ (see above).

1270 **REFERENCES**

1271 1. Colautti, R. I., Maron, J. L. & Barrett, S. C. H. Common garden comparisons of native and
1272 introduced plant populations: latitudinal clines can obscure evolutionary inferences. *Evol.*
1273 *Appl.* **2**, 187–199 (2009).
1274 2. Oduor, A. M. O., Leimu, R. & van Kleunen, M. Invasive plant species are locally adapted
1275 just as frequently and at least as strongly as native plant species. *J. Ecol.* **104**, 957–968
1276 (2016).
1277 3. Hodgins, K. A., Bock, D. G. & Rieseberg, L. H. Trait evolution in invasive species. *Annual*
1278 *Plant Reviews Online* 459–496 (2018).
1279 4. Colautti, R. I. & Barrett, S. C. H. Rapid adaptation to climate facilitates range expansion of
1280 an invasive plant. *Science* **342**, 364–366 (2013).
1281 5. Essl, F. *et al.* Biological flora of the British isles:Ambrosia artemisiifolia. *J. Ecol.* **103**, 1069–
1282 1098 (2015).
1283 6. Bagarozzi, D. A., Jr & Travis, J. Ragweed pollen proteolytic enzymes: possible roles in
1284 allergies and asthma. *Phytochemistry* **47**, 593–598 (1998).
1285 7. Schaffner, U. *et al.* Biological weed control to relieve millions from Ambrosia allergies in
1286 Europe. *Nat. Commun.* **11**, 1745 (2020).
1287 8. Lake, I. R. *et al.* Climate Change and Future Pollen Allergy in Europe. *Environ. Health*
1288 *Perspect.* **125**, 385–391 (2017).
1289 9. Ziska, L. *et al.* Recent warming by latitude associated with increased length of ragweed
1290 pollen season in central North America. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4248–4251
1291 (2011).
1292 10. van Boheemen, L. A. & Hodgins, K. A. Rapid repeatable phenotypic and genomic
1293 adaptation following multiple introductions. *Mol. Ecol.* **29**, 4102–4117 (2020).
1294 11. Bieker, V. C. *et al.* Uncovering the hologenomic basis of an extraordinary plant invasion.
1295 *bioRxiv* (2022) doi:10.1101/2022.02.03.478494.

1296 12. Chauvel, B., Dessaint, F., Cardinal-Legrand, C. & Bretagnolle, F. The historical spread of
1297 Ambrosia artemisiifolia L. in France from herbarium records. *J. Biogeogr.* **33**, 665–673
1298 (2006).
1299 13. van Boheemen, L. A. *et al.* Multiple introductions, admixture and bridgehead invasion
1300 characterize the introduction history of Ambrosia artemisiifolia in Europe and Australia. *Mol.*
1301 *Ecol.* **26**, 5421–5434 (2017).
1302 14. van Boheemen, L. A., Atwater, D. Z. & Hodgins, K. A. Rapid and repeated local adaptation
1303 to climate in an invasive plant. *New Phytol.* **222**, 614–627 (2019).
1304 15. Seebens, H. *et al.* No saturation in the accumulation of alien species worldwide. *Nat.*
1305 *Commun.* **8**, 14435 (2017).
1306 16. Fisher, R. A. *The Genetical Theory of Natural Selection*. (The Clarendon Press, 1930).
1307 17. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, 1983).
1308 18. Orr, H. A. The population genetics of adaptation: The distribution of factors fixed during
1309 adaptive evolution. *Evolution* **52**, 935–949 (1998).
1310 19. Yeaman, S. & Whitlock, M. C. The genetic architecture of adaptation under migration-
1311 selection balance. *Evolution* **65**, 1897–1911 (2011).
1312 20. Yeaman, S., Gerstein, A. C., Hodgins, K. A. & Whitlock, M. C. Quantifying how constraints
1313 limit the diversity of viable routes to adaptation. *PLoS Genet.* **14**, e1007717 (2018).
1314 21. Hermisson, J. & Pennings, P. S. Soft sweeps: molecular population genetics of adaptation
1315 from standing genetic variation. *Genetics* **169**, 2335–2352 (2005).
1316 22. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation.
1317 *Genetics* **173**, 419–434 (2006).
1318 23. Wellenreuther, M. & Bernatchez, L. Eco-Evolutionary Genomics of Chromosomal
1319 Inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
1320 24. Krimbas, C. B. & Powell, J. R. *Drosophila Inversion Polymorphism*. (CRC Press, 1992).
1321 25. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes

1322 to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**,
1323 (2010).
1324 26. Todesco, M. *et al.* Massive haplotypes underlie ecotypic differentiation in sunflowers.
1325 *Nature* **584**, 602–607 (2020).
1326 27. Kirkpatrick, M. & Barrett, B. Chromosome inversions, adaptive cassettes and the evolution
1327 of species' ranges. *Mol. Ecol.* **24**, 2046–2055 (2015).
1328 28. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome
1329 assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
1330 29. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for
1331 long-range linkage. *Genome Res.* **26**, 342–350 (2016).
1332 30. Payne, W. W., Raven, P. H. & Kyhos, D. W. Chromosome numbers in compositae. Iv.
1333 Ambrosieae. *Am. J. Bot.* **51**, 419–424 (1964).
1334 31. Kubešová, M. *et al.* Naturalized plants have smaller genomes than their non-invading
1335 relatives: a flow cytometric analysis of the Czech alien flora. *Preslia* **82**, 81–96 (2010).
1336 32. Bai, C., Alverson, W. S., Follansbee, A. & Waller, D. M. New reports of nuclear DNA
1337 content for 407 vascular plant taxa from the United States. *Ann. Bot.* **110**, 1623–1629
1338 (2012).
1339 33. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update:
1340 Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage
1341 for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654
1342 (2021).
1343 34. Barker, M. S. *et al.* Multiple paleopolyploidizations during the evolution of the Compositae
1344 reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.*
1345 **25**, 2445–2455 (2008).
1346 35. Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering
1347 and Asterid evolution. *Nature* **546**, 148–152 (2017).

1348 36. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database
1349 management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491
1350 (2011).
1351 37. Zagotta, M. T. *et al.* The Arabidopsis ELF3 gene regulates vegetative photomorphogenesis
1352 and the photoperiodic induction of flowering. *Plant J.* **10**, 691–702 (1996).
1353 38. Martin, A. & Orgogozo, V. The Loci of repeated evolution: a catalog of genetic hotspots of
1354 phenotypic variation. *Evolution* **67**, 1235–1250 (2013).
1355 39. Gautier, M. Genome-Wide Scan for Adaptive Divergence and Association with Population-
1356 Specific Covariates. *Genetics* **201**, 1555–1579 (2015).
1357 40. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for
1358 global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
1359 41. Booker, T. R., Yeaman, S. & Whitlock, M. C. The WZA: A window-based method for
1360 characterizing genotype-environment association. *bioRxiv* 2021.06.25.449972 (2021)
1361 doi:10.1101/2021.06.25.449972.
1362 42. Prapas, D. *et al.* QTL analysis reveals an oligogenic architecture of a rapidly adapting trait
1363 during the European invasion of common ragweed. *bioRxiv* (2022)
1364 doi:10.1101/2022.02.24.481758.
1365 43. Werck-Reichhart, D., Bak, S. & Paquette, S. Cytochromes p450. *Arabidopsis Book* **1**,
1366 e0028 (2002).
1367 44. Gupta, R. & Chakrabarty, S. K. Gibberellic acid in plant: still a mystery unresolved. *Plant*
1368 *Signal. Behav.* **8**, (2013).
1369 45. Li, H. & Ralph, P. Local PCA Shows How the Effect of Population Structure Differs Along
1370 the Genome. *Genetics* **211**, 289–304 (2019).
1371 46. Haldane, J. B. S. The theory of a cline. *J. Genet.* **48**, 277–284 (1948).
1372 47. Thurman, T. J. & Barrett, R. D. H. The genetic consequences of selection in natural
1373 populations. *Mol. Ecol.* **25**, 1429–1448 (2016).

1374 48. Chen, K.-W., Marusciac, L., Tamas, P. T., Valenta, R. & Panaitescu, C. Ragweed Pollen
1375 Allergy: Burden, Characteristics, and Management of an Imported Allergen Source in
1376 Europe. *Int. Arch. Allergy Immunol.* **176**, 163–180 (2018).
1377 49. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764
1378 (2013).
1379 50. Sun, Y. *et al.* Climate warming can reduce biocontrol efficacy and promote plant invasion
1380 due to both genetic and transient metabolomic changes. *Ecol. Lett.* **25**, 1387–1400 (2022).
1381 51. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
1382 assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
1383 52. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
1384 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
1385 53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1386 *arXiv [q-bio.GN]* (2013).
1387 54. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology
1388 Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
1389 55. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming,
1390 identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
1391 56. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and
1392 phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056–1082
1393 (2014).
1394 57. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK,*
1395 *and WDL in Terra*. ('O'Reilly Media, Inc.', 2020).
1396 58. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-
1397 generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
1398 59. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158
1399 (2011).

1400 60. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-
1401 Generation Reference Panels. *The American Journal of Human Genetics* vol. 103 338–348
1402 (2018).
1403 61. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable
1404 element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
1405 62. Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality
1406 control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).
1407 63. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. (2015).
1408 64. Acids research, N. & 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic*
1409 *Acids Res.* **49**, D480–D489 (2021).
1410 65. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in
1411 eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
1412 66. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
1413 67. Berardini, T. Z. *et al.* The arabidopsis information resource: Making and mining the 'gold
1414 standard' annotated reference plant genome. *genesis* vol. 53 474–485 (2015).
1415 68. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1416 (2009).
1417 69. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor*
1418 *Improv* **27**, 1–26 (2009).
1419 70. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of
1420 flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Res.* **44**, D1167–71
1421 (2016).
1422 71. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
1423 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;
1424 iso-2; iso-3. **6**, 80–92 (2012).
1425 72. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer

1426 datasets. *GigaScience* vol. 4 (2015).
1427 73. Hijmans, R. J. *et al.* Package ‘raster’. *R package* **734**, (2015).
1428 74. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).
1429 75. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-
1430 wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
1431 76. Fox, J. *et al.* The car package. *R Foundation for Statistical Computing* **1109**, (2007).
1432 77. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation
1433 Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).
1434 78. Meisner, J. & Albrechtsen, A. Inferring Population Structure and Admixture Proportions in
1435 Low-Depth NGS Data. *Genetics* **210**, 719–731 (2018).
1436 79. Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. Package ‘emmeans’. (2019).
1437 80. Rastas, P., Paulin, L., Hanski, I., Lehtonen, R. & Auvinen, P. Lep-MAP: fast and accurate
1438 linkage map construction for large SNP datasets. *Bioinformatics* **29**, 3128–3134 (2013).
1439 81. Otto, S. P. & Day, T. *A Biologist’s Guide to Mathematical Modeling in Ecology and*
1440 *Evolution*. (Princeton University Press, 2011).
1441 82. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics*. (Sinauer, 2007).
1442 83. Connallon, T. & Clark, A. G. Antagonistic versus nonantagonistic models of balancing
1443 selection: characterizing the relative timescales and hitchhiking effects of partial selective
1444 sweeps. *Evolution* **67**, 908–917 (2013).
1445 84. Chevin, L., Gompert, Z. & Nosil, P. Frequency dependence and the predictability of
1446 evolution in a changing environment. *Evolution Letters* vol. 6 21–33 (2022).
1447 85. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory*. vol. 26 977 (The
1448 Blackburn Press, 1971).
1449 86. Felsenstein, J. *Theoretical Evolutionary Genetics*. (2019).
1450 87. Roughgarden, J. *Theory of population genetics and evolutionary ecology: an introduction*.
1451 (sidalc.net, 1979).

1452 88. Lasne, C., Sgrò, C. M. & Connallon, T. The Relative Contributions of the X Chromosome
1453 and Autosomes to Local Adaptation. *Genetics* **205**, 1285–1304 (2017).
1454 89. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).