Title

Large haploblocks underlie rapid adaptation in an invasive weed

One Sentence Summary

Parallel evolution between native and invasive ranges of *Ambrosia artemisiifolia* is aided by putative chromosomal inversions.

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*‡]

* Corresponding author

[†] These authors contributed equally

[‡] These authors contributed equally

Affiliations

[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 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 between ranges (18%), are associated with rapidly adapting traits, and exhibit dramatic frequency shifts over space and time. These results highlight the importance of structural and large-effect variants in rapid adaptation, which have been critical to *A. artemisiifolia*'s global spread.

Main text

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, observable over human lifespans, as invasive species must often adapt to environments that drastically differ from those of native ranges. 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 also inhabit broad and climatically diverse ranges, which favours the evolution of adaptations to local environmental conditions (*1*, *2*), along with evolved tolerance of environmental extremes, which may be conducive to invasiveness (*3*). Because they 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 result in 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 (common ragweed), an annual weed native to North America, has mounted successful invasions on all continents except Antarctica (*4*). 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 (*5*). In Europe alone, approximately 13.5 million people suffer from *Ambrosia*-induced allergies, costing ~7.4 billion euros annually (*6*). Continued invasion and climate change are predicted to more than double sensitization to *Ambrosia* pollen (*7*), 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 (*8*). 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 connectivity among populations (*9*, *10*). In Europe, invasive populations have been established through multiple genetically diverse introductions from North America over the last ~160 years (*10*–*12*). Remarkably, latitudinal clines observed for multiple traits in the native range, including flowering time and size, have re-evolved in Europe (as well as Australia; *13*), suggesting rapid local adaptation following invasion. Moreover, this trait-level parallelism is echoed in signals of parallelism at the genetic level (*9*).

As biological invasions continue to increase worldwide (*14*) and the effects of anthropogenic climate change intensify, understanding the genetic architecture of adaptation to extreme environmental shifts—a classical question in evolutionary research—becomes ever more important. While long-standing theory suggests that evolution in response to incremental change should proceed through mutations of infinitesimally small (*15*) or moderate (*16*) effect,

large-effect mutations are predicted to be useful in bridging extreme, sudden environmental shifts (*17*). 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 (*18*). Large-effect mutations are also more likely to produce patterns of evolutionary repeatability, or genetic parallelism, between species' ranges (*19*), 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 (*20*). These features of large-effect mutations may, however, be achieved by groups of mutations in tight genetic linkage (*18*), including those mutations captured by chromosomal inversions (*21*). There is substantial empirical evidence for the involvement of inversions in local adaptation (*22*). For example, *Drosophila melanogaster*'s *In(3R)Payne* inversion shows parallel environmental associations across multiple continents (*23*), and several plant inversions have been identified as contributing to local adaptation and ecotype divergence (*24*, *25*). Theory also predicts that inversions can drive range expansions (*26*), though their actual contributions to biological invasions are not well-understood.

Here we describe a new, near chromosome-level reference *A. artemisiifolia* genome assembly with alignments of over 600 modern and historic genome sequences from throughout North America and Europe (*10*). These spatially and temporarily resolved data provide unparalleled insight into this intriguing model of rapid adaptation, by enhancing our ability to detect selection and estimate its strength, providing information about the tempo of adaptive evolutionary change, and allowing the extent of parallel genetic adaptation to be uncovered through comparisons of native and invasive ranges. Our genomic data also allow for the identification of genome-structural variants, and we demonstrate for the first time the importance of multiple large-effect standing structural variants for adaptation in an invasive plant.

Parallel patterns in spatial signatures of local adaptation

To identify genome-wide spatial signatures of local adaptation in North American and European ranges, we performed genome scans for population allele frequencies among *A. artemisiifolia* modern samples that were both highly divergent between populations (BayPass XtX; *27*), and correlated with environmental variables (latitude, longitude and 19 WorldClim temperature and precipitation variables; table S1; *28*). Statistics were analyzed in 10kb windows using the Weighted-Z Analysis (WZA; *29*). In North America (143 samples; 43 populations) 1,480 (84.0%) of the 1,762 outlier windows for genomic divergence (XtX) were also outlier windows for at least one environmental variable (XtX-EAA), while in Europe (141 samples; 31 populations), only 908 (51.7%) of the 1,755 XtX outlier windows overlapped environmental variable outlier windows. Signatures of local adaptation were much stronger in North America than Europe, with the North American range showing more extreme XtX values (fig. 1C), as well as more XtX-EAA windows (table S1; fig. 1C). This is consistent with the expectation that samples from the native range will be better-adapted to their environment than those from the recently-invaded European range.

GWAS using 121 individuals across North American (n = 43) and European (n = 78) ranges identified significant associations with 13 of 30 putatively adaptive phenotypes previously measured by (*13*; fig. S1; table S2). Six phenotypes yielded associations within predicted genes, including an association between flowering time phenotypes and a nonsynonymous SNP in the *A. artemisiifolia* homologue of *A. thaliana* flowering-time pathway gene *early flowering 3* (*ELF3*) on scaffold 2 (table S2; fig. 1A, B). Candidate SNPs in *ELF3* are restricted to high-latitude populations in both ranges, where they occur at moderate to high frequencies (fig. 1C), indicative of parallel local adaptation between ranges.

Previous studies in *A. artemisiifolia* have identified signatures of repeatability between native and invasive ranges at phenotypic and genetic levels (*9*, *13*). We observed congruent patterns

in our data: among North American and European XtX-EAA outlier windows, 173 showed parallel associations with the same environmental variable between ranges (significantly more than would be expected by chance; hypergeometric $p = 4.39\times10^{-63}$; fig. 1C). To account for the possibility that this pattern is driven by extended linkage disequilibrium at a smaller number of loci, we combined consecutive windows, and windows in haploblock regions, into single windows and repeated the analysis, in which the parallelism remained highly significant (hypergeometric $p = 1.08\times10^{-63}$). Consequently, many of the same regions of the genome are involved in climate adaptation in both ranges.

In addition to *ELF3*, parallel XtX-EAA outlier windows included four windows within two previously identified flowering-onset QTL (*30*). North American, European and parallel XtX-EAA outlier windows included 17, ten, and three flowering-time pathway genes respectively, however, this only represented a significant enrichment (Fisher's exact test; $p < 0.05$) in North America (table S3). Gene ontology terms enriched in parallel XtX-EAA windows included "iron ion binding" and "heme binding" (terms relating to cytochrome P450 genes), as well as "gibberellin biosynthetic process" (table S4). Some cytochrome P450 genes are involved in detoxification of xenobiotic compounds and the synthesis of defense compounds, while others play key developmental roles (*31*), including contributing the biosynthesis of gibberellin, a hormone that regulates a range of developmental events, including flowering (*32*).

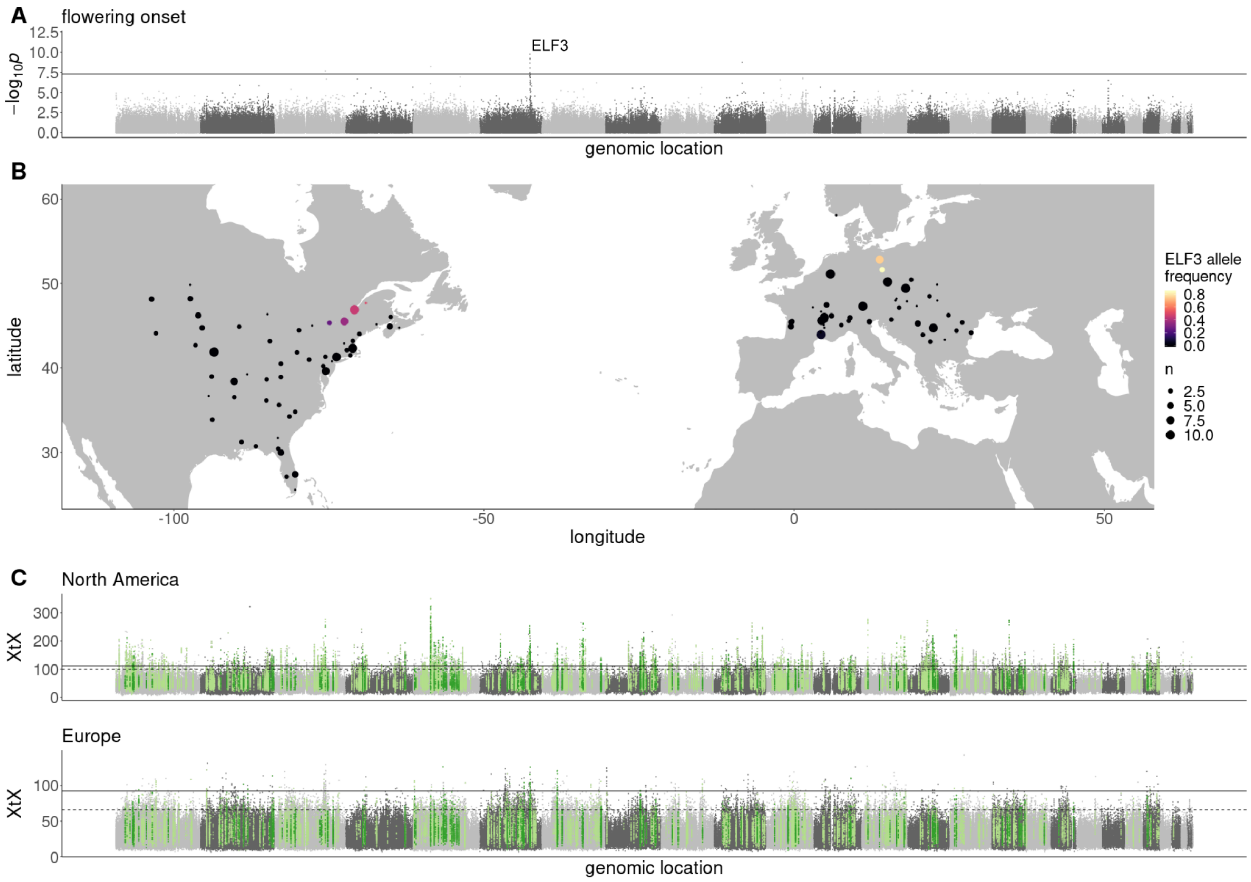

Figure 1. Signatures of climate adaptation in *Ambrosia artemisiifolia.* **A.** GWAS results (-$\log_{10}p$ against genomic location) for flowering onset (solid line indicates a Bonferroni-corrected p-value of 0.05). **B.** Distribution of a strongly-associated nonsynonymous SNP in *ELF3* among modern *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.

Temporal evidence of rapid adaptation during invasion

To further identify the features of recent adaptation in *A. artemisiifolia*, we leveraged herbarium samples, collected from as early as 1830, for phenotypic and genomic analyses. A trait-based analysis of digitized herbarium images (fig. S2) identified a significant shift in flowering and fruiting over time in Europe (fig. S3; table S5). For the trait presence of a mature male inflorescence, we identified a significant interaction between collection year and latitude ($F_{1,886}$ =

7.89, $p < 0.01$) where in northern populations, more recently collected plants were more likely to be flowering than older historic specimens. Similar patterns were identified with the presence of fruit, as older samples were less likely to produce fruit later in the season compared to recent samples (day-by-year interaction $F_{1,886} = 32.33$, $p < 0.001$; fig. S3; table S6). This pattern is consistent with experimental common garden evidence that earlier flowering has evolved in northern populations and later flowering in southern populations following the invasion of Europe (*13*). Further, this shift in both flowering and fruit set over time supports the hypothesis that a mismatch between the local environment and the genotypes present impacted the reproductive output of *A. artemisiifolia* in its initial stages of colonization, particularly in northern Europe.

Genome sequences of herbarium samples (*10*) allowed comparisons between historic and modern population pairs (seven in North America and ten in Europe), which were scanned for signatures of local selective sweeps by identifying windows with extreme shifts in allele frequency and extreme reductions in diversity over time. We found far more evidence for recent sweeps in Europe (654 unique windows) than in North America (119 unique windows; fig. S4 c.f. S5; table S7), consistent with the expectation that a haphazardly-introduced invader will frequently be maladapted initially and undergo rapid adaptation to local environmental conditions following its introduction. The most dramatic selective sweep signatures were observed in Berlin and multiple French populations in a ~14Mb region on scaffold 448 (fig. 2A; fig. S5), which accounts for 374 (79%) of the European sweep windows. In Berlin, sweep windows were also observed containing and surrounding the flowering onset GWAS peak that includes *ELF3*. Furthermore, the early-flowering GWAS candidate in *ELF3* increased over time in populations within a 200km radius of Berlin (4.9% to 73.9%), with an allele frequency shift greater than all putatively neutral loci sampled over the same time period (fig. 2C); similar shifts were not observed in any of the North American populations with comparably high frequencies

of the early flowering *ELF3* allele (fig. S6). In comparisons of spatial and temporal signatures of selection, sweep windows are also XtX-EAA outlier windows far more often in Europe (6.4%) than in North America (0.8%; table S7), consistent with rapid local adaptation in the recently-established European range, contrasting more gradual, ongoing local adaptation in the native North American range. Among gene annotations in North American sweeps, there were no flowering-time genes identified, while in Europe, ten flowering-time genes were within sweep windows (representing a significant enrichment; Fisher's exact test $p = 0.027$; table S3) as well as a dramatic enrichment of the gene ontology term "recognition of pollen" (table S8). These enrichments are attributable to the sweep windows on scaffold 448, as both flowering-time and recognition of pollen enrichments are no longer enriched after removing windows on this scaffold.

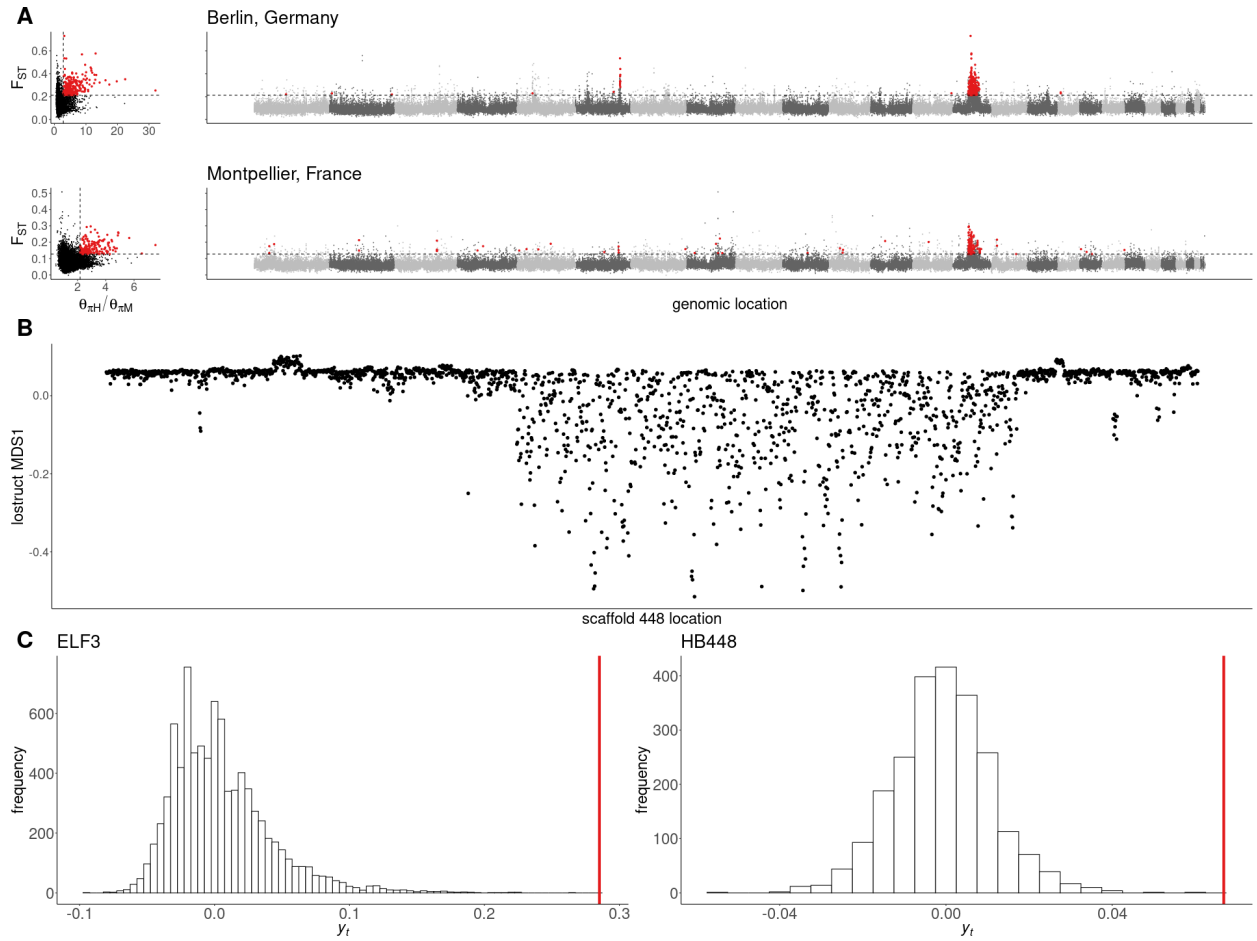

Figure 2. Temporal signatures of selective sweeps in Europe. A. Distributions of F_{ST} between historic and modern samples and the ratio of historic to modern nucleotide diversity ($\theta_{\pi H}/\theta_{\pi M}$) from Berlin and Montpellier, and F_{ST} against genomic location. Red points indicate 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 SNPs sampled from northern populations (left) and y_t calculated from the haplotype frequency of HB448 (red line), against a distribution of y_t values from 7,394 SNPs sampled from across the European range (right).

Haploblocks underlying parallel regions of local adaptation

Previously, chromosomal inversions have been identified as driving local adaptation of ecotypes in species of *Helianthus* (sunflower; *25*). We used a similar approach to identify signatures of putative inversions (haploblocks) contributing to local adaptation in *A. artemisiifolia* (fig. S7). Focussing our analysis on regions showing signatures of adaptation, we identified seven haploblocks overlapping the 173 WZA windows that were parallel outliers for both XtX and at least one climate variable. These haploblocks contained 31 of the 173 windows (17.9%; fig. S8), with one haploblock also corresponding to the selective sweep region on scaffold 448 (fig. 2). This suggests these haploblocks have played a pivotal role in generating parallel signatures of selection observed in *A. artemisiifolia*. Additionally, ABBA-BABA tests (*33*) across the scaffolds harboring these haploblocks provide evidence that some haploblocks may be the result of introgression from related *Ambrosia* species (fig. S9). Six haploblocks showed evidence of genotype-specific reductions in recombination using multiple genetic maps of each region, which is the pattern expected when recombination is suppressed by inversions ((*30*); fig. S10).

To identify changes in haploblock frequency over time and space, which would be consistent with selection on these putative inversions, we genotyped all historic and modern samples by local, divergent population structure in haploblock regions (fig. S7) and ran generalized linear models comparing haplotype frequency with latitude, time (historic vs. modern) and range (North America vs. Europe; fig. 3A; fig. S11; table S9-12). Significant associations with latitude in one or more range or timepoints were identified in five of the seven haploblocks (HB5, HB21, HB27a, HB27b and HB31), indicative of climate adaptation. In HB27b and HB31, we identified significant three-way interactions between latitude, time and range (fig. 3A). In HB27b haplotype frequency was correlated with latitude in North America at both timepoints, as well as in modern Europe, consistent with climate-mediated selection. However, historic European

populations did not display an association with latitude, which may reflect maladaptation in Europe during the initial stages of range expansion. HB21 and HB5 displayed parallel associations with latitude in modern and historic North America and Europe, although for HB21 the slope was greater for modern populations in both ranges, potentially indicative of temporal changes in selection in the native range. For modern samples, we also compared the slopes of these five haploblocks with a genome-wide distribution of slopes for North America and Europe using 10,000 unlinked SNPs randomly selected from outside haploblocks and genes. All but HB5 fell into the 5% tail of the distribution for at least one of the ranges (fig. S12).

We used spatial variation in haploblock frequency to estimate the strength of selection on these haploblocks using contemporary data (see supplementary text S2). We specifically compared estimates of the maximum slope of each putative inversion frequency cline to simple population-genetic models for clines at equilibrium between local selection and gene flow. In these models, cline slopes are proportional to $\sqrt{s}/\sigma$, where s represents the fitness effect of a given inversion and σ is the average dispersal distance of individuals in a given range ((*34*); supplementary text S2). While our estimates of selection are, therefore, scaled by the dispersal rate, dispersal should equally affect all inversions within a given range, allowing us to infer the relative strength of spatially varying selection for each putative inversion. First, estimates of relative selection were correlated between the ranges indicating parallel patterns of selection along the latitudinal gradient across haploblocks (r = 0.79, p = 0.03). We also find that the change in slope from historic to modern European populations was significantly correlated with our estimates of the relative strength of spatially varying selection in Europe on these seven haploblocks (r = 0.97, p = 0.0001; fig. S13). This would be expected if stronger selection caused greater change in haplotype frequency with latitude if European populations were initially far from the local optimum. The same pattern was not identified in the North American native range, whose historic populations should be closer to equilibrium (r = 0.25, p = 0.58). For haploblocks

exhibiting significant latitudinal clines in at least one time point or range, historic slopes were also significantly shallower than modern slopes in Europe ($t_{4.58}$ = 4.58, p = 0.015; mean absolute slope: historic EU = 0.05; modern EU = 0.20), but not in North America ($t_{6.61}$ = -0.58, p = 0.58), 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. The haploblock was not associated with latitude, but it showed a significant time and range effect, with a significantly higher frequency in modern Europe compared to North America (historic and modern) and historic Europe. Signatures of selective sweeps were observed in the HB448 region in multiple European populations (fig. 2A; fig. S5). 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 7,394 SNPs matched for allele frequency (fig. 2C). None of the loci in the 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 2.5% difference in fitness (95% CI = {1.9%, 3.2%}) 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 (see supplementary text S1). This estimate is smaller than many empirical estimates of selection on individual loci in natural populations (*35*). The greater timespan of our sampling facilitated detection of strong signals of selection for loci that would otherwise be missed by the short time periods afforded by most temporal studies.

Several analyses provide evidence for the biological function of these adaptive haploblocks. From genome annotations, we found that, as a group, haploblocks were collectively enriched for

flowering-time pathway genes (Fisher's exact test $p = 0.034$), although individual haploblocks were not (table S13). HB448 was enriched for the "recognition of pollen" gene ontology term; 28 (17%) of the genome's 167 genes annotated with this term fall within this haploblock (table S14). HB27a was enriched for genes with the "pectate lyase activity" term, including the top BLAST hit for *Amba1* (99.7% identity, *E*-value = 0), which encodes the *A. artemisiifolia* protein responsible for the majority of allergic reactions (*36*). A detailed analysis of the HB27a region reveals a cluster of six closely-related pectate lyase genes which correspond with elevated XtX and XtX-EAA outlier windows in both ranges (fig. S14). HB27b corresponds to a large-effect flowering time and height QTL (*30*). HB2 also overlaps the flowering time GWAS candidate *ELF3* (fig. 1A;B), and the nonsynonymous variant which displays strong patterns in GWAS is only observed on one of the haploblock genotype backgrounds. We also identified phenotypic associations with haploblocks by encoding haploblock genotypes into our GWAS pipeline. Significant associations ($p < 0.05$ Bonferroni-corrected for multiple testing across seven haploblocks) were observed for five haploblocks including several for HB21 (including traits related to flowering time, size, and senescence), as well as HB27b (end of flowering), HB448 (stem diameter) and HB31 (total biomass; fig. 3B; table S15). Given that overdominance is often implicated as the selective force maintaining inversion polymorphisms (*22*), we tested for the effect of overdominance in haploblock alleles on phenotypes. We identified significant differences in dichogamy between individuals with heterozygous and homozygous genotypes for both HB31 and HB448 (fig. 3B; table S15).

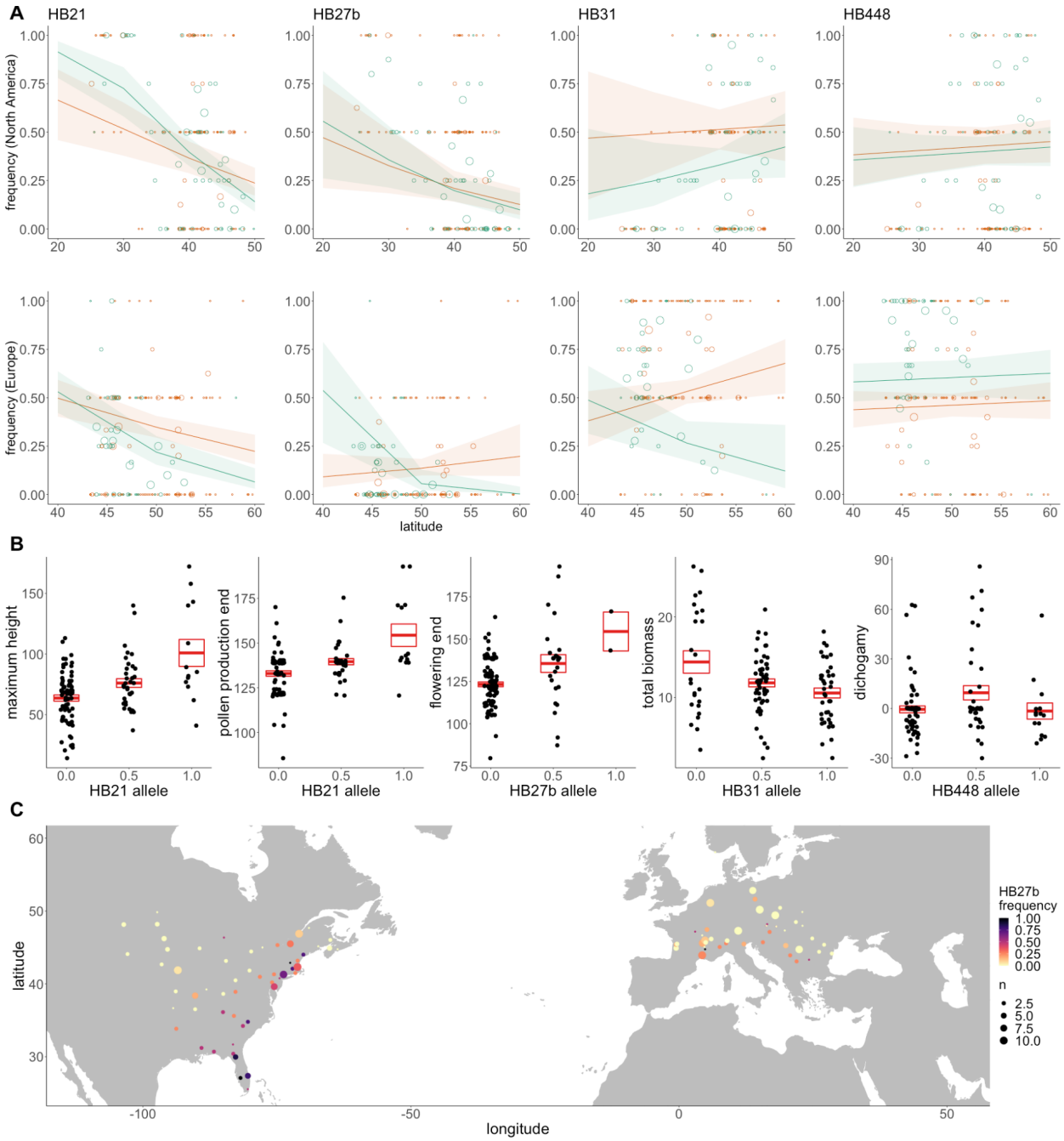

Figure 3. Haploblock distributions and trait associations. A. Frequency of four haploblocks against latitude of sampling for historic (brown) and modern (green) *A. artemisiifolia* populations (circles; scaled relative to population sample size). Lines represent reduced logistic regression models with 95% CI ribbons (see table S9-12 for model details). **B.** Examples of significant associations between haploblock alleles and phenotypes (boxes denote mean and SEM). **C.** HB27b allele frequency in modern *A. artemisiifolia* populations.

Conclusions

We have described, at an unprecedented temporal and spatial resolution, the evolutionary-genetic changes accompanying a recent and rapid invasion by a noxious pest. Our study system, while unique in many ways, yields results with important general implications for our understanding of the genetic basis of rapid adaptation to environmental change and the pervasiveness of parallel evolution in geographically widespread species. Through multiple introductions over the last ~160 years, a large proportion of the standing genetic variation in *A. artemisiifolia*'s native range has been co-opted for adaptation in and across the European invasive range, resulting in striking patterns of parallel evolution between North America and Europe. Parallel evolution is a hallmark of natural selection and parallel changes at the genetic level point to constraints and biases in the genetic pathways to adaptation that are evolutionarily achievable; when paths to adaptation are limited, evolution will repeatedly use the same small subset of genes to reach the same adaptive endpoints. From herbarium specimens that were sampled throughout the time course of *A. artemisiifolia*'s invasion, we observed an abrupt change in flowering and fruiting over time, and strong temporal signatures of selection in regions of the genome associated with flowering. These results provide compelling evidence for rapid local adaptation in European populations, with the strongest genetic signals of rapid change over time typically conferring the strongest signatures of local adaptation in our spatial analyses.

Large haploblocks (putative inversions) contribute substantially to these genetic signals of parallel adaptation. We propose that these haploblocks maintain cassettes of co-selected genes that effectively segregate as single alleles of large effect (*21*, *26*), providing a genetic architecture suited to local adaptation in the face of high gene flow (*10*, *18*). Consistent with this hypothesis, haploblocks are enriched for genes with particular biological functions, display associations with locally-adaptive traits, and carry signals of strong selection in both the native and invasive ranges. The evolution of inversions along environmental gradients has been

reported in a range of species (*22*). However, by investigating haploblocks in an invasive plant with extensive timestamped collections, we have demonstrated dramatic and adaptive evolutionary change of putative inversions under natural conditions, providing compelling evidence of strong and recent natural selection. These data have also allowed us to directly estimate selection for these variants, and we have shown that haploblocks with the strongest estimates of clinal selection are driven more rapidly towards their putative equilibria within the invasive range.

An important question during this era of environmental upheaval is the role of adaptation during range expansion and its necessity during colonization. Through our analysis of historic samples, we have shown that *A. artemisiifolia* was present in regions throughout Europe well before many of these adaptive variants became locally common, suggesting the species' extensive plasticity may have facilitated its initial expansion. Strong local selection further improved the match between genotypes and local environments, even appearing to affect reproductive output in herbarium specimens. Many of the selected variants we identified are linked to traits (e.g. days to flowering onset, days to the end of pollen production, and biomass) that are key factors in the timing, length and severity of the local pollen season. Consequently, contemporary local adaptation has played a central role in shaping the allergy season in Europe and will likely continue to be critical as climate change and continued range expansion further amplify the damaging effects of this hazardous weed.

References

1. K. A. Hodgins, D. G. Bock, L. H. Rieseberg, Trait evolution in invasive species. *Annual Plant Reviews Online*, 459–496 (2018).

2. A. M. O. Oduor, R. Leimu, M. van Kleunen, Invasive plant species are locally adapted just as frequently and at least as strongly as native plant species. *J. Ecol.* **104**, 957–968 (2016).

3. R. I. Colautti, S. C. H. Barrett, Rapid adaptation to climate facilitates range expansion of an invasive plant. *Science*. **342**, 364–366 (2013).

4. F. Essl, K. Biró, D. Brandes, O. Broennimann, J. M. Bullock, D. S. Chapman, B. Chauvel, S. Dullinger, B. Fumanal, A. Guisan, G. Karrer, G. Kazinczi, C. Kueffer, B. Laitung, C. Lavoie, M. Leitner, T. Mang, D. Moser, H. Müller-Schärer, B. Petitpierre, R. Richter, U. Schaffner, M. Smith, U. Starfinger, R. Vautard, G. Vogl, M. von der Lippe, S. Follak, Biological flora of the British isles: *Ambrosia artemisiifolia*. *J. Ecol.* **103**, 1069–1098 (2015).

5. D. A. Bagarozzi Jr, J. Travis, Ragweed pollen proteolytic enzymes: possible roles in allergies and asthma. *Phytochemistry*. **47**, 593–598 (1998).

6. U. Schaffner, S. Steinbach, Y. Sun, C. A. Skjøth, L. A. de Weger, S. T. Lommen, B. A. Augustinus, M. Bonini, G. Karrer, B. Šikoparija, M. Thibaudon, H. Müller-Schärer, Biological weed control to relieve millions from *Ambrosia* allergies in Europe. *Nat. Commun.* **11**, 1745 (2020).

7. I. R. Lake, N. R. Jones, M. Agnew, C. M. Goodess, F. Giorgi, L. Hamaoui-Laguel, M. A. Semenov, F. Solmon, J. Storkey, R. Vautard, M. M. Epstein, Climate Change and Future Pollen Allergy in Europe. *Environ. Health Perspect.* **125**, 385–391 (2017).

8. L. Ziska, K. Knowlton, C. Rogers, D. Dalan, N. Tierney, M. A. Elder, W. Filley, J. Shropshire, L. B. Ford, C. Hedberg, P. Fleetwood, K. T. Hovanky, T. Kavanaugh, G. Fulford, R. F. Vrtis, J. A. Patz, J. Portnoy, F. Coates, L. Bielory, D. Frenz, Recent warming by latitude associated with increased length of ragweed pollen season in central North America. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4248–4251 (2011).

9. L. A. van Boheemen, K. A. Hodgins, Rapid repeatable phenotypic and genomic adaptation following multiple introductions. *Mol. Ecol.* **29**, 4102–4117 (2020).

10. V. C. Bieker, P. Battlay, B. Petersen, X. Sun, J. Wilson, J. C. Brealey, F. Bretagnolle, K. Nurkowski, C. Lee, G. L. Owens, J. Y. Lee, F. L. Kellner, L. van Boheeman, S. Gopalakrishnan, M. Gaudeul, H. Mueller-Schaerer, G. Karrer, B. Chauvel, Y. Sun, L. Dalén, P. Poczai, L. H. Rieseberg, M. T. P. Gilbert, K. A. Hodgins, M. D. Martin, Uncovering the hologenomic basis of an extraordinary plant invasion. *bioRxiv* (2022), doi:10.1101/2022.02.03.478494.

11. B. Chauvel, F. Dessaint, C. Cardinal-Legrand, F. Bretagnolle, The historical spread of *Ambrosia artemisiifolia* L. in France from herbarium records. *J. Biogeogr.* **33**, 665–673 (2006).

12. L. A. van Boheemen, E. Lombaert, K. A. Nurkowski, B. Gauffre, L. H. Rieseberg, K. A. Hodgins, Multiple introductions, admixture and bridgehead invasion characterize the introduction history of *Ambrosia artemisiifolia* in Europe and Australia. *Mol. Ecol.* **26**, 5421–5434 (2017).

13. L. A. van Boheemen, D. Z. Atwater, K. A. Hodgins, Rapid and repeated local adaptation to climate in an invasive plant. *New Phytol.* **222**, 614–627 (2019).

14. H. Seebens, T. M. Blackburn, E. E. Dyer, P. Genovesi, P. E. Hulme, J. M. Jeschke, S. Pagad, P. Pyšek, M. Winter, M. Arianoutsou, S. Bacher, B. Blasius, G. Brundu, C. Capinha, L. Celesti-Grapow, W. Dawson, S. Dullinger, N. Fuentes, H. Jäger, J. Kartesz, M. Kenis, H. Kreft, I. Kühn, B. Lenzner, A. Liebhold, A. Mosena, D. Moser, M. Nishino, D. Pearman, J. Pergl, W. Rabitsch, J. Rojas-Sandoval, A. Roques, S. Rorke, S. Rossinelli, H. E. Roy, R. Scalera, S. Schindler, K. Štajerová, B. Tokarska-Guzik, M. van Kleunen, K. Walker, P. Weigelt, T. Yamanaka, F. Essl, No saturation in the accumulation of alien species worldwide. *Nat. Commun.* **8**, 14435 (2017).

15. R. A. Fisher, *The Genetical Theory of Natural Selection* (The Clarendon Press, 1930).

16. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).

17. H. A. Orr, The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. *Evolution*. **52**, 935–949 (1998).

18. S. Yeaman, M. C. Whitlock, The genetic architecture of adaptation under migration-selection balance. *Evolution*. **65**, 1897–1911 (2011).

19. S. Yeaman, A. C. Gerstein, K. A. Hodgins, M. C. Whitlock, Quantifying how constraints limit the diversity of viable routes to adaptation. *PLoS Genet.* **14**, e1007717 (2018).

20. J. Hermisson, P. S. Pennings, Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*. **169**, 2335–2352 (2005).

21. M. Kirkpatrick, N. Barton, Chromosome inversions, local adaptation and speciation. *Genetics*. **173**, 419–434 (2006).

22. M. Wellenreuther, L. Bernatchez, Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).

23. C. B. Krimbas, J. R. Powell, *Drosophila Inversion Polymorphism* (CRC Press, 1992).

24. D. B. Lowry, J. H. Willis, A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8** (2010), doi:10.1371/journal.pbio.1000500.

25. M. Todesco, G. L. Owens, N. Bercovich, J.-S. Légaré, S. Soudi, D. O. Burge, K. Huang, K. L. Ostevik, E. B. M. Drummond, I. Imerovski, K. Lande, M. A. Pascual-Robles, M. Nanavati, M. Jahani, W. Cheung, S. E. Staton, S. Muños, R. Nielsen, L. A. Donovan, J. M. Burke, S. Yeaman, L. H. Rieseberg, Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*. **584**, 602–607 (2020).

26. M. Kirkpatrick, B. Barrett, Chromosome inversions, adaptive cassettes and the evolution of species' ranges. *Mol. Ecol.* **24**, 2046–2055 (2015).

27. M. Gautier, Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*. **201**, 1555–1579 (2015).

28. S. E. Fick, R. J. Hijmans, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).

29. T. R. Booker, S. Yeaman, M. C. Whitlock, The WZA: A window-based method for characterizing genotype-environment association. *bioRxiv* (2021), p. 2021.06.25.449972.

30. D. Prapas, R. Scalone, J. Lee, K. Nurkowski, S. Bou-assi, L. H. Rieseberg, P. Battlay, K. A. Hodgins, QTL analysis reveals an oligogenic architecture of a rapidly adapting trait during the European invasion of common ragweed. *bioRxiv* (2022),

doi:10.1101/2022.02.24.481758.

31. D. Werck-Reichhart, S. Bak, S. Paquette, Cytochromes p450. *Arabidopsis Book*. **1**, e0028 (2002).

32. R. Gupta, S. K. Chakrabarty, Gibberellic acid in plant: still a mystery unresolved. *Plant Signal. Behav.* **8** (2013), doi:10.4161/psb.25504.

33. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).

34. J. B. S. Haldane, The theory of a cline. *J. Genet.* **48**, 277–284 (1948).

35. T. J. Thurman, R. D. H. Barrett, The genetic consequences of selection in natural populations. *Mol. Ecol.* **25**, 1429–1448 (2016).

36. K.-W. Chen, L. Marusciac, P. T. Tamas, R. Valenta, C. Panaitescu, Ragweed pollen allergy: burden, characteristics, and management of an imported allergen source in Europe. *Int. Arch. Allergy Immunol.* **176**, 163–180 (2018).

37. M. Kubešová, L. Moravcova, J. Suda, V. Jarošík, P. Pyšek, Others, Naturalized plants have smaller genomes than their non-invading relatives: a flow cytometric analysis of the Czech alien flora. *Preslia*. **82**, 81–96 (2010).

38. C. Bai, W. S. Alverson, A. Follansbee, D. M. Waller, New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Ann. Bot.* **110**, 1623–1629 (2012).

39. D. Guan, S. A. McCarthy, J. Wood, K. Howe, Y. Wang, R. Durbin, Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. **36**, 2896–2898 (2020).

40. M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, E. M. Zdobnov, BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

41. N. H. Putnam, B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette, R. Calef, C. J. Troll, A. Fields, P. D. Hartley, C. W. Sugnet, D. Haussler, D. S. Rokhsar, R. E. Green, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).

42. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

43. M. Schubert, L. Ermini, C. Der Sarkissian, H. Jónsson, A. Ginolhac, R. Schaefer, M. D. Martin, R. Fernández, M. Kircher, M. McCue, E. Willerslev, L. Orlando, Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056–1082 (2014).

44. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes*. **9**, 88 (2016).

45. G. A. Van der Auwera, B. D. O'Connor, *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* ("O'Reilly Media, Inc.," 2020).

46. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

47. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics*. **27**, 2156–2158 (2011).

48. B. L. Browning, Y. Zhou, S. R. Browning, A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*. **103** (2018), pp.

338–348.

49. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

50. M. S. Campbell, M. Law, C. Holt, J. C. Stein, G. D. Moghe, D. E. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C. J. Lawrence, D. Ware, S.-H. Shiu, K. L. Childs, Y. Sun, N. Jiang, M. Yandell, MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).

51. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. 2013--2015 (2015).

52. C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. **12**, 491 (2011).

53. N. Acids research, 2021, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

54. G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. **23**, 1061–1067 (2007).

55. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics*. **5**, 59 (2004).

56. T. Z. Berardini, L. Reiser, D. Li, Y. Mezheritsky, R. Muller, E. Strait, E. Huala, The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *genesis*. **53** (2015), pp. 474–485.

57. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).

58. A. Alexa, J. Rahnenführer, Gene set enrichment analysis with topGO. *Bioconductor Improv*. **27**, 1–26 (2009).

59. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* . **6**, 80–92 (2012).

60. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. **4** (2015), , doi:10.1186/s13742-015-0047-8.

61. R. J. Hijmans, J. Van Etten, J. Cheng, M. Mattiuzzi, M. Sumner, J. A. Greenberg, O. P. Lamigueiro, A. Bevan, E. B. Racine, A. Shortridge, Others, Package "raster." *R package*. **734** (2015) (available at ftp://h64-50-233-100.mdsnwi.tisp.static.tds.net/pub/cran/web/packages/raster/raster.pdf).

62. R. C. Team, Others, R: A language and environment for statistical computing (2013) (available at http://r.meteo.uni.wroc.pl/web/packages/dplR/vignettes/intro-dplR.pdf).

63. H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, E. Eskin, Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

64. J. Fox, G. G. Friendly, S. Graves, R. Heiberger, G. Monette, H. Nilsson, B. Ripley, S. Weisberg, M. J. Fox, M. Suggests, The car package. *R Foundation for Statistical Computing*. **1109** (2007) (available at http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/packages/car.pdf).

65. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. **15**, 356 (2014).

66. H. Li, P. Ralph, Local PCA shows how the effect of population structure differs along the genome. *Genetics*. **211**, 289–304 (2019).

67. J. Meisner, A. Albrechtsen, Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*. **210**, 719–731 (2018).

68. R. Lenth, H. Singmann, J. Love, P. Buerkner, M. Herve, Package "emmeans" (2019), (available at https://cran.microsoft.com/snapshot/2018-01-13/web/packages/emmeans/emmeans.pdf).

69. P. Rastas, L. Paulin, I. Hanski, R. Lehtonen, P. Auvinen, Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics*. **29**, 3128–3134 (2013).

70. S. P. Otto, T. Day, *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*

(Princeton University Press, 2011).

71. D. L. Hartl, A. G. Clark, *Principles of Population Genetics* (Sinauer, 2007).

72. T. Connallon, A. G. Clark, Antagonistic versus nonantagonistic models of balancing selection: characterizing the relative timescales and hitchhiking effects of partial selective sweeps. *Evolution*. **67**, 908–917 (2013).

73. L. Chevin, Z. Gompert, P. Nosil, Frequency dependence and the predictability of evolution in a changing environment. *Evolution Letters*. **6** (2022), pp. 21–33.

74. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (The Blackburn Press, 1971), vol. 26.

75. J. Felsenstein, *Theoretical Evolutionary Genetics* (2019).

76. J. Roughgarden, *Theory of Population Genetics and Evolutionary Ecology: An Introduction* (sidalc.net, 1979).

77. C. Lasne, C. M. Sgrò, T. Connallon, The relative contributions of the X chromosome and autosomes to local adaptation. *Genetics*. **205**, 1285–1304 (2017).

78. S. Wright, Evolution in Mendelian populations. *Genetics*. **16**, 97–159 (1931).

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 the National Infrastructure for High Performance Computing and Data Storage in Norway (UNINETT Sigma2). We thank the curators from the following herbaria for allowing us to destructively sample their 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., the Australian Research Council (grants DP220102362 and DP180102531) and Human Frontier Science Program (grant 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 **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

Competing interests

The authors declare no competing interests

Data and materials availability

DNA sequences and computer code will be made available through NCBI and GitHub.

Supplementary materials

Materials and Methods

Fig. S1-S19

Text S1-S3

Table S1-S19 (separate XLSX file)

References (37-78)

Materials and 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 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 ~1.2-Gbp (*37*, *38*). As the sequenced individual was heterozygous, we expected that large heterozygous contigs (haplotigs) were incorrectly assembled as distinct contigs, resulting in a falsely inflated assembly. In order to remove one haplotig from each heterozygous pair, we used Purge_dups (*39*). Purge_dups uses both read depth and sequence homology to identify and remove duplicated contigs or duplicated regions of a contig, which reduced our assembly size to 1.2-Gbp (table S16). BUSCO (Benchmarking Universal Single-Copy Orthologs; *40*) comparisons of original and reduced assemblies (table S16) demonstrated a 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 (*41*). Dovetail OmniC library sequences were aligned to the draft input assembly using bwa (*42*). 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. S15). HiRise scaffolding

resulted in a final assembly consisting of 1,215,680,099bp, 451 scaffolds (L_{90} = 20), 253 complete BUSCO genes (99.2%; 138 single copy [54.1%]; 115 duplicated [45.1%]), no fragmented BUSCO genes and two missing BUSCO genes (0.8%; table S16).

Sample alignment, variant calling and filtering

FASTQ files from historic and modern *A. artemisiifolia* samples from North America and Europe (*10*) were aligned to our new reference genome using the Paleomix pipeline (*43*), which incorporates AdapterRemoval (*44*), BWA-MEM (*42*), Picard MarkDuplicates (https://broadinstitute.github.io/picard/) and GATK IndelRealigner (*45*). For modern samples, variants were called using GATK UnifiedGenotyper (*46*); GATK VariantFiltration (*45*) and VcfTools (*47*) were used to filter variant calls. SNP and indel calls were separately filtered using GATK hard-filtering recommendations (SNPs: QD < 2.0, FS > 60.0, SOR > 3.0, MQ < 40.0, ReadPosRankSum < -8.0, MQRankSum < -12.5; indels: QD < 2.0, FS > 200.0, SOR > 10.0, ReadPosRankSum < -20.0, InbreedingCoeff < -0.8). Additionally, SNPs and indels were separately filtered for sites with depth (DP) less than one standard deviation below the mean, and greater than 1.5 standard deviations above the mean. Individual genotypes were set to missing if their depth was less than three, then variants with greater than 20% missing across all samples were removed. Samples with greater than 50% missing variants were removed. For the remaining 311 modern samples, contigs with fewer than 100 variants were removed, and the remaining 30 contigs were phased and imputed using Beagle 5.2 (*48*).

Genome annotation

Prior to annotation of the genome, repetitive elements were identified using RepeatModeler2 (*49*). ProExcluder (*50*) was then run to remove any protein coding genes from the repeat library. Finally, RepeatMasker (*51*) was used to mask the genome using the finalized repeat library (table S17).

To obtain RNA transcript sequences for annotation of the genome, after 160 days of growth additional samples of leaf, stem, flower, root, and branch were taken from individual “NSS02/B” and flash-frozen in liquid nitrogen. From these we extracted RNA from seven tissues (young leaf, old leaf, stem, branch, and three stages of development of the floral head) using a Spectrum Plant Total RNA Kit (Sigma, USA) with on-column DNA digestion following the manufacturer's protocol. RNA extracts from all five tissues were pooled into a single sample. mRNA was enriched using oligo (dT) beads, and the first strand cDNA was synthesized using the Clontech SMARTer PCR cDNA Synthesis Kit, followed by first-strand synthesis with SMARTScribeTM Reverse Transcriptase. After cDNA amplification, a portion of the product was used directly as a non-size selected SMRTbell library. In parallel, the rest of amplification was first selected using either BluePippin or SageELF, and then used to construct a size-selected SMRTbell library after size fractionation. DNA damage and ends were then repaired, followed by hairpin adaptor ligation. Finally, sequencing primers and polymerase were annealed to SMRTbell templates, and IsoSeq isoform sequencing was performed by Novogene Europe (Cambridge, UK) using a PacBio Sequel II instrument, yielding 97,819,215 HiFi reads. To prepare the raw IsoSeq RNA data for downstream use in the annotation of the genome, we first identified the transcripts in the PacBio single-molecule sequencing data by following the IsoSeq v3 pipeline provided by PacificBiosciences (https://github.com/PacificBiosciences/IsoSeq). Briefly, the pipeline takes PacBio subread files as an input and undergoes steps of consensus generation, demultiplexing of primers, IsoSeq3 refinement, followed by a final clustering of the reads.

Genome annotation was performed using the genome annotation pipeline MAKER2 version 2.31.9 (*52*) with *ab initio* and homology-based gene predictions (fig. S16). 3,819 unique UniProtKB/Swiss-Prot protein sequences (*53*) from asterids (a monophyletic group of flowering

plants), *Asteraceae* (sunflower family), and *Ambrosia* (ragweeds) were used for homology-based gene prediction. As no training gene models were available for *A. artemisiifolia*, we used CEGMA (*54*) to train the *ab initio* gene predictor SNAP (*55*). MAKER2 was run with command-line arguments *model_org=simple*, *softmask=1*, *augustus_species=arabidopsis* and the *snaphmm* parameter was set to the HMM generated in the manual training of SNAP. As expressed sequence tag (EST) evidence, we used the IsoSeq clustered reads merged with a previously described transcriptome (*10*).

29,849 predicted *A. artemisiifolia* proteins (table S18) were compared with *Arabidopsis thaliana* annotations (TAIR10 representative gene model proteins (*56*); table S19) using the *blastp* command in BLAST+(*57*). 26,688 predicted *A. artemisiifolia* proteins (89.4%) matched *A. thaliana* genes with an E-value < 1×10^{-6}; these annotations were retained for downstream analysis. Gene ontology (GO) enrichment was assessed using GO terms from *A. thaliana* TAIR 10 (*56*) BLAST results. To identify GO terms enriched among candidate lists, the R/topGO package (*58*) was used with Fisher's exact test, the 'weight01' algorithm, and a p-value < 0.05 to assess significance. Additionally, annotations were cross-referenced with 306 *A. thaliana* FLOR-ID flowering time pathway genes (Bouché *et al.* 2016). 538 predicted *A. artemisiifolia* genes were matched to this dataset, representing 212 unique FLOR-ID genes. Enrichment of flowering time genes was also assessed in candidate gene lists using Fisher's exact test and a $p < 0.05$ threshold. The effects of imputed variants on predicted genes were estimated using SnpEff (*59*).

Allele frequency outliers and environmental allele associations

Imputed genotype data from modern samples were divided for between-range and within-range analyses in PLINK 1.9 (*60*), and a minor allele frequency threshold of 0.05 was applied within data subsets. For within-range analyses, sampling locations with fewer than two samples were

excluded and allele frequencies were calculated for each sampling location, resulting in 955,827 SNPs across 143 samples and 43 populations in North America and 941,649 SNPs across 141 samples and 31 populations for Europe. Allele frequency outliers were identified within each range using the BayPass core model (*27*), with an Ω covariance matrix computed from 10,000 randomly-sampled SNPs that were located outside annotated genes and haploblocks, and pruned for linkage disequilibrium using a window size of 50kb, a step size of 5bp and an r^2 of 0.5 in PLINK (*60*). To identify allele frequency variation associated with environmental variables within ranges, 19 bioclimatic variables were extracted for each sampling location from the WorldClim database (*28*) using the R/raster package (*61*). Associations with latitude and longitude of sampling location, along with the 19 bioclimatic variables, were assessed for correlations with population allele frequencies using Kendall's τ statistic in R (*62*). Genome-wide XtX and τ results were analyzed in non-overlapping 10kb windows using the windowed-Z analysis (WZA; *29*), with the top 5% of windows designated outliers.

Genome-wide association studies

Imputed genotypes from modern samples were filtered in PLINK 1.9 (*60*). Non-SNP sites and sites with more than two alleles were removed. The 121 samples overlapping those phenotyped by (*13*) were retained, and sites with a minor allele frequency below 0.05 were removed, resulting in 950,742 SNPs for analysis. Genome-wide association studies (GWAS) were performed across 121 individuals from both North American (n = 43) and European (n = 78) ranges using EMMAX (*63*), and incorporating an identity-by-state kinship matrix (generated in PLINK 1.9; *60*) to account for genetic structure among samples. The kinship matrix was computed using 650,301 SNPs which remained after pruning for linkage disequilibrium using a window size of 50kb, a step size of 5bp and an r^2 of 0.5. Candidate SNPs were identified using a conservative threshold of Bonferroni-corrected p-values < 0.05.

Phenotypic analysis of herbarium specimens

We conducted a trait-based analysis of herbarium specimens found in the Global Biodiversity Information Facility database (gbif.org 2021). We compiled information from all *A. artemisiifolia* European herbarium specimens for which there was a digitized image of the individual in the database alongside corresponding metadata (location and collection date). The collection date spanned 1849 to 2020 (median 1975) and comprised 985 specimens. We determined the stage of flowering (no male inflorescence present, only immature male inflorescence present, mature male inflorescence present) for each image. The presence of fruit was also recorded. The male inflorescence was used as an indicator of flowering as these structures are more visually prominent than female flowers and the onset of male and female flowering is highly correlated (*13*). Male florets consist of prominent spike-like racemes of male capitula, and are found at the terminus of the stem, whereas female florets are observed to be in inconspicuous cyme-like clusters and are arranged in groups at the axils of main and lateral stem leaves (fig. S2). The dates when the specimens were collected were converted to Julian day of the year. We conducted a generalized linear model with a binomial response and logit link (glm R). Both binary traits (presence of a mature male inflorescence; the presence of fruit) were included as response variables in two separate models. The significance of the effects were tested using the Anova function (car package R; *64*) using type 3 tests. For both models, the predictors of latitude, 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 (*13*) and we expected northern populations to evolve early flowering to match the shorter growing seasons in these areas. Conversely, late flowering was expected to evolve in southern environments due to a trade-off between size and timing of reproduction.

Historic-modern genomic comparisons

To identify targets of recent selection, we compared historic and modern population samples from 16 locations (seven from North America and nine from Europe; table S7). Analyses were performed in ANGSD (*65*) 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 the spatial analysis of modern populations, and 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 European populations within 200km of Berlin, whilst the frequency of HB448 in Europe was ascertained in using the haploblock frequency estimates below. To understand the magnitude of these allele frequency shifts relative to putatively neutral alleles elsewhere in the genome, we calculated a standardized measure of frequency change,

y_t, using estimates of historic, p_0, and contemporary, p_t, allele frequencies according to the equation:

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

where *t* is the number of generations separating the frequency estimates (equivalent to the number of years due to ragweed's annual lifecycle). As we show in supplementary text S3, the distribution of y_t estimates under neutrality are predictable and roughly independent of the initial frequency of each neutral variant once the loci with low-frequency initial minor allele frequencies are filtered out. To further assess if selection was the likely cause of temporal changes of the *ELF3* and HB448 variants, we estimated the distribution of y_t estimates for computed from 10,000 randomly-sampled SNPs that were located outside annotated genes and haploblocks and pruned for linkage disequilibrium using a window size of 50kb, a step size of 5bp and an r^2 of 0.5 in PLINK (*60*). Prior to calculation of y_t, sampled SNPs were then filtered for a minor allele frequency > 0.2 for HB448 comparisons and MAF > 0.05 for *ELF3* comparisons (due to the low historic frequency of *ELF3* in historic Berlin populations). We then compared the distributions to the y_t values of candidate adaptation loci to test whether candidate regions were more divergent than the putatively neutral distribution. As a point of comparison we repeated this analysis for the *ELF3* allele in Quebec. As in Berlin, this allele is at high frequencies, but substantial temporal change was not expected as the populations were predicted to be closer to the equilibrium over the temporal sampling period in the native range. Samples designated as belonging to Quebec populations were pooled at both timepoints. Allele frequency changes of the 10,000 randomly-sampled SNPs and the non-synonymous *ELF3* allele were assessed as above.

Haploblock identification

To identify signatures of large, segregating haploblocks across the genome, we performed local windowed principal component analysis with Lostruct (*66*). Using SNP data from all modern samples, we extracted the first ten multidimensional scaling (MDS) coordinates across each of the 30 largest scaffolds in windows of 100 SNPs. These MDS coordinates were then plotted along each scaffold to observe regions of local structure, indicative of segregating haploblocks. We focused on outlier MDS signals that overlapped parallel outlier windows for both XtX and at least one environmental variable, and also showed well-defined boundaries indicative of chromosomal inversions. We tested for additional evidence of inversions using PCA of MDS outlier regions and kmeans clustering in R (*62*) to identify regions containing three distinct clusters representing heterozygotes and two homozygotes, and furthermore tested that putative inversion heterozygotes showed higher average heterozygosity across the region compared with putative homozygotes.

Haploblock frequency changes over time and and space

For seven candidate inversions, a local PCA of each region and kmeans clustering was then repeated in PCAngsd (*67*), so as to allow genotyping of all historic and modern samples for the haploblock haplotypes. We also conducted a PCA on 10,000 randomly-sampled SNPs that were located outside annotated genes and haploblocks, and pruned for linkage disequilibrium using a window size of 50 kb, a step size of 5 bp and an r^2 of 0.5 in PLINK (*60*). Following this, we used generalized linear models (glm R) to assess how haplotype frequency (binomial response with logit link) changed over time and space. A count of each haplotype at a geographic location was the binomial response variable and time period (historic or modern), range (North America or Europe), latitude, and all interactions between these three main effects were used as predictors. Non-significant interactions were removed in a stepwise fashion,

starting with the highest order. PC1 and PC2 were included as covariates to control for the effects of population structure on haplotype frequency. We tested the significance of the effects in our model using the Anova function (car package R; *64*) with type 3 tests. Significant differences among groups for means or slopes were tested with the emmeans package using an fdr correction (*68*).

We estimated the relative strength of selection on haploblocks along the latitudinal clines in modern North American and European populations using slopes from logistic regressions (see supplementary text S2). Specifically, we used generalized linear models to estimate the slopes of the regression for each range and time point combination (group). As above, a count of each haplotype at a geographic location was the binomial response variable and time period (historic or modern), range (North America or Europe), latitude, and all interactions between these three main effects were used as predictors. All interactions were retained in the model and slopes and their confidence intervals estimated for each group using the function emtrends (emmeans package R (*68*); table S12). PC1 and PC2 were included as covariates to control for the effects of population structure on haplotype frequency. We expected the slopes to be shallower in the historic versus the modern European group, but similar across timepoints in North America. To test this, we used a t-test and compared slopes for modern and historic timepoints in each range for those haploblocks showing significant latitudinal clines in at least one group. We also expected that the change of the slope over time would be the greatest in haploblocks showing the largest estimates of selection in Europe (table S12). We estimated the relative strength of selection for the modern European range for each haploblock and tested if the change in slope for each haploblock was correlated with this estimate. We also examined if there was a correlation in the relative strength of selection for modern North American and European haploblocks, which would indicate parallel selection along the cline in each range.

We compared our slope estimates of the haploblocks to the genome wide distribution in each range using 10,000 randomly selected SNPs outside of annotated genes and haploblocks. We did this to determine if our haploblocks showed stronger latitudinal patterns than a genome wide sample of SNPs, which may be indicative of spatially varying selection. For the modern samples in each range (North America or Europe), we fit a generalized linear model with latitude as the only predictor. We did this for each SNP and each haploblock that was statistically associated with latitude.

Tests of introgression in haploblock regions

Genome sequences from *Ambrosia trifida* (giant ragweed), *Ambrosia psilostachya* (western ragweed) and *Ambrosia carduacea* (*10*) were aligned to our *A. artemisiifolia* reference genome using the steps previously described for *A. artemisiifolia* samples. To assess the evidence for introgression contributing to haploblock formation, we performed ABBA-BABA tests (*33*) in 10Mb sliding windows with a 1Mb step size across each scaffold harboring a haploblock of interest. Analyses were run in ANGSD, using eight samples homozygous for each haploblock allele as H1 and H2 respectively, and *A. carduacea* as the outgroup (H4). Analyses were run with either five samples of *A. trifida* or three samples of *A. psilostachya* as H3.

Recombination rates in haploblocks

To determine if there were genotype-specific reductions in recombination rate in the seven candidate haploblocks, which would be consistent with inversions, we made use of three previously generated genetic maps (*30*). Markers were generated using genotype by sequencing and alignments to the new reference. Details of the sequencing, alignments and variant calling can be found in (*30*). We developed sex-specific genetic maps using Lep-MAP3 (*69*) for each scaffold of interest and in each mapping population (an F1 mapping population and two F2 mapping populations). Multiple maps were constructed since the haploblocks may

have been segregating in different frequencies in the parents of the mapping populations, which were derived from outcrossing. Linkage map construction was constrained by the physical order of the markers along each scaffold of interest. Genetic distance (cM) was plotted against physical position along the chromosome for each map and the intervals of the flowering time and height QTL (*30*) on scaffold 27 as well as the boundaries of the candidate haploblocks were visualized and inspected for reduced recombination compared to the rest of the scaffold.

Supplementary Figures

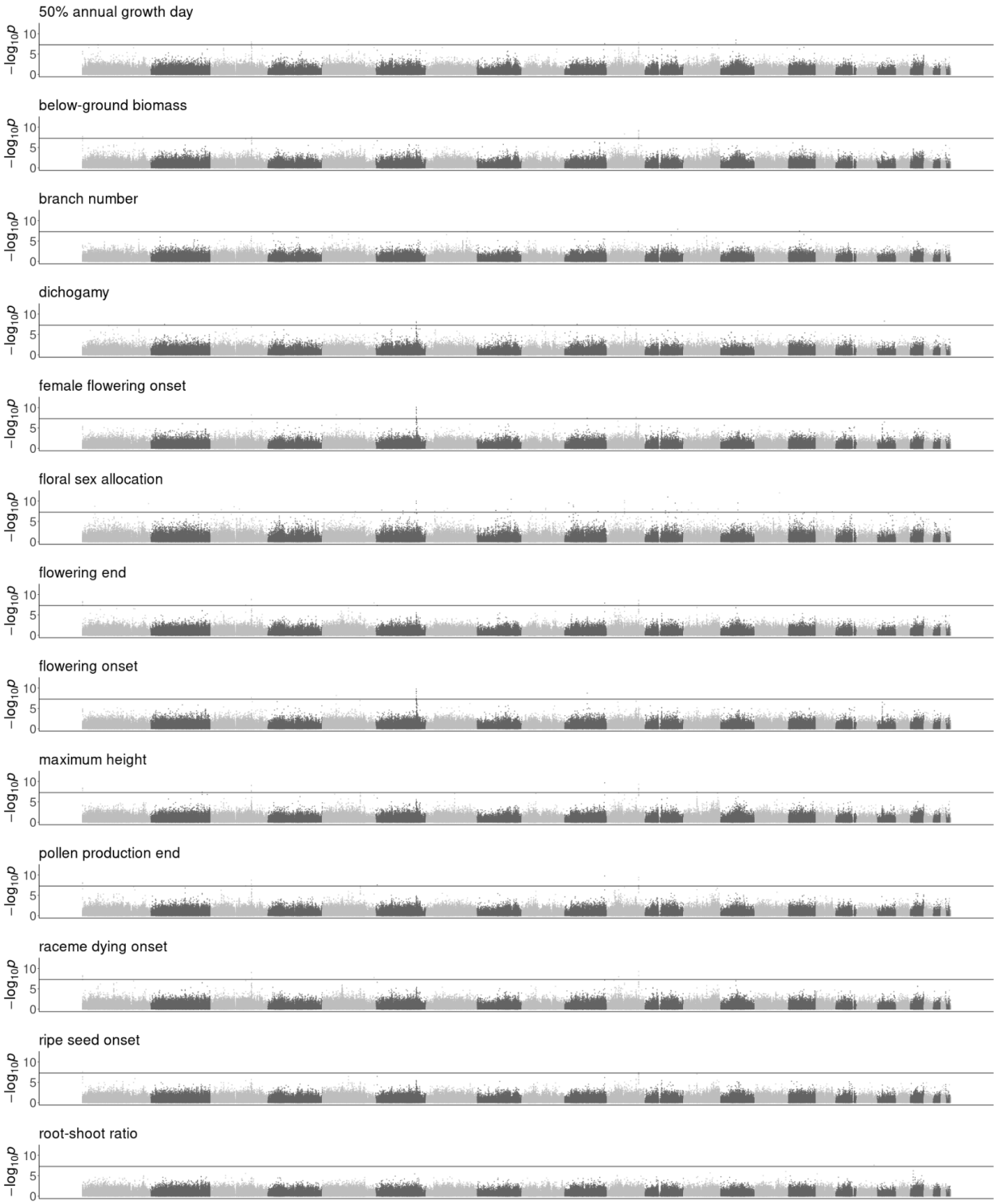

Figure S1. Genome-wide association study results for phenotypes with significant SNPs. Solid lines indicate a Bonferroni-corrected significance threshold of 0.05.

Figure S2. An example herbarium specimen of *Ambrosia artemisiifolia* (left) and a detail (right) showing the mature male inflorescence (solid line) and seeds (dashed lines).

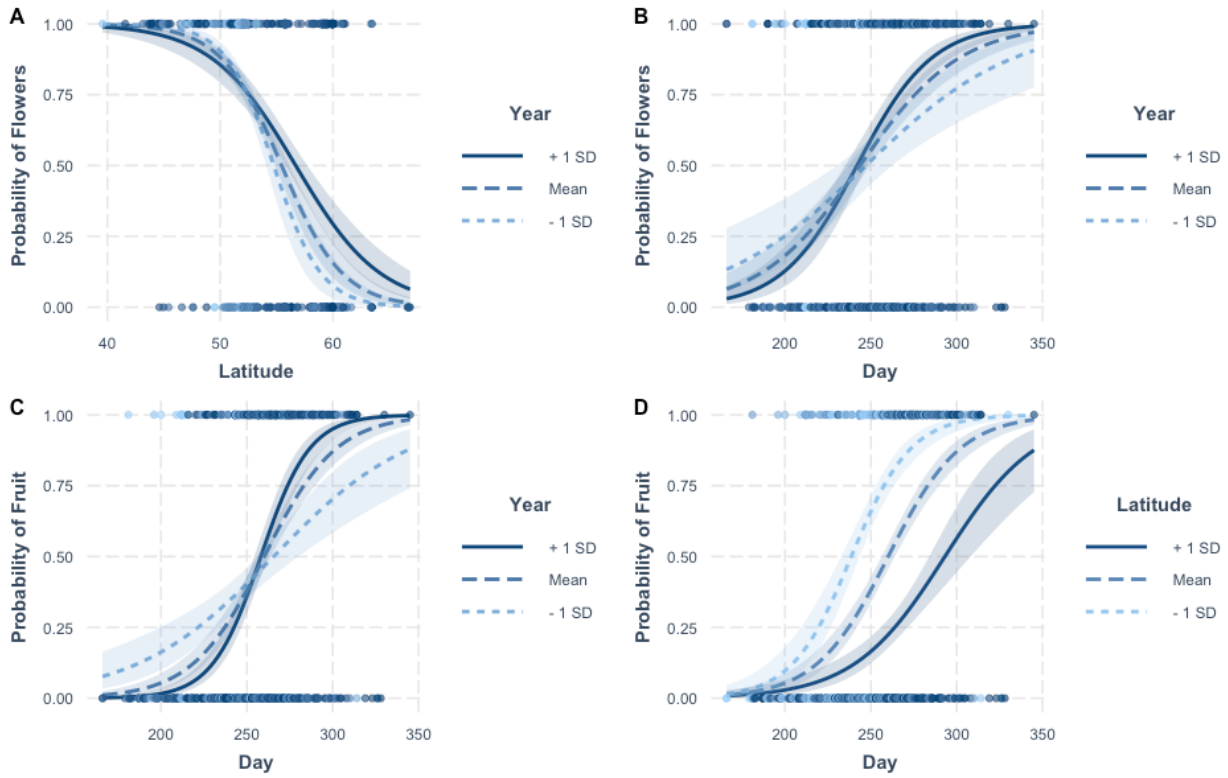

Figure S3. Interaction plots illustrating the results of generalized linear models examining the presence of mature male inflorescences (probability of flowers) or mature fruit (probability of fruit) in herbarium specimens of *A. artemisiifolia* in Europe as a function of collection day (Day), latitude 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.

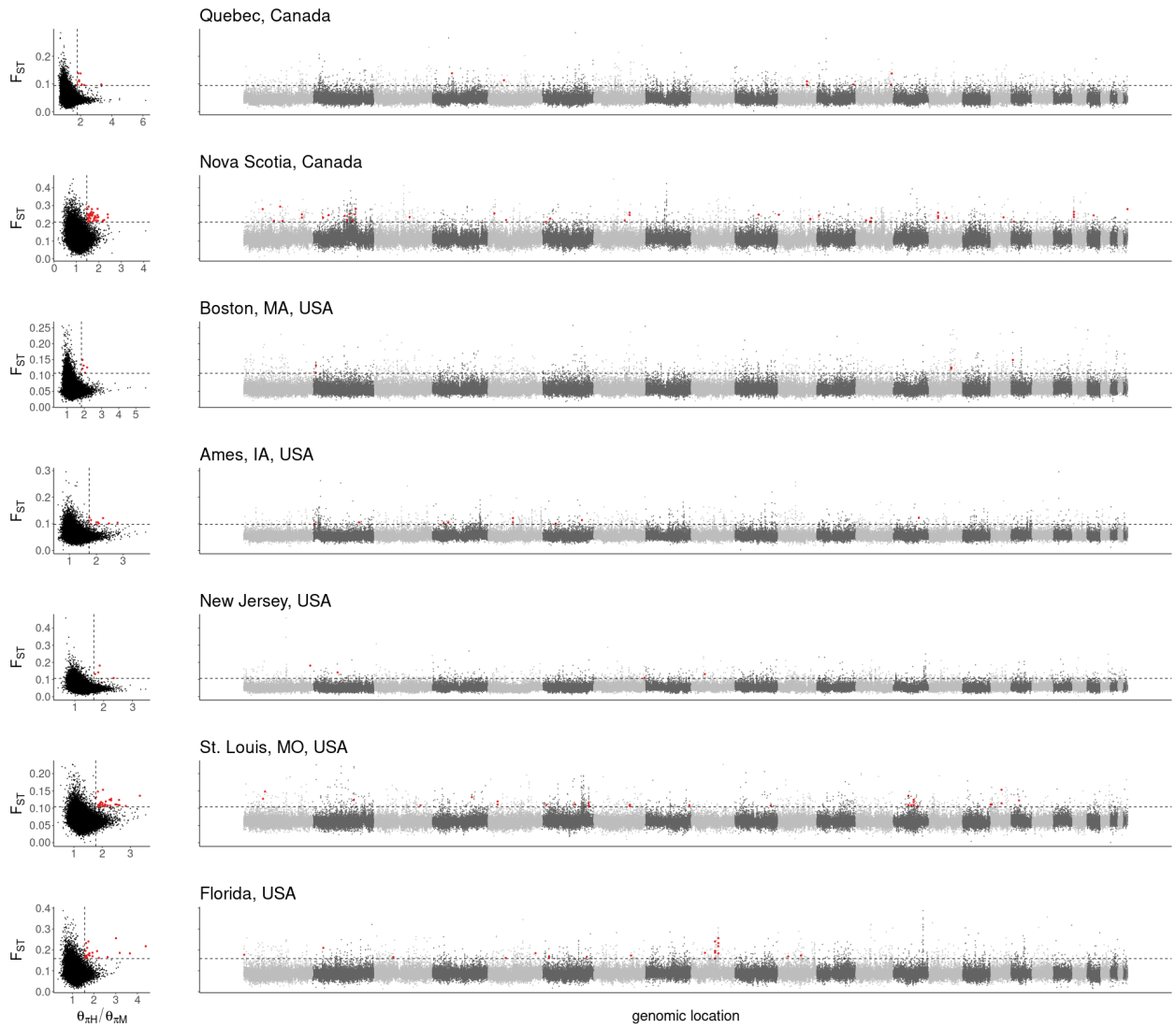

Figure S4. Distributions of F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ between historic and modern samples from North American populations, and F_{ST} against genomic location. Red points indicate putative selective sweep windows, which are in top one percent of per-window F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ (dashed lines).

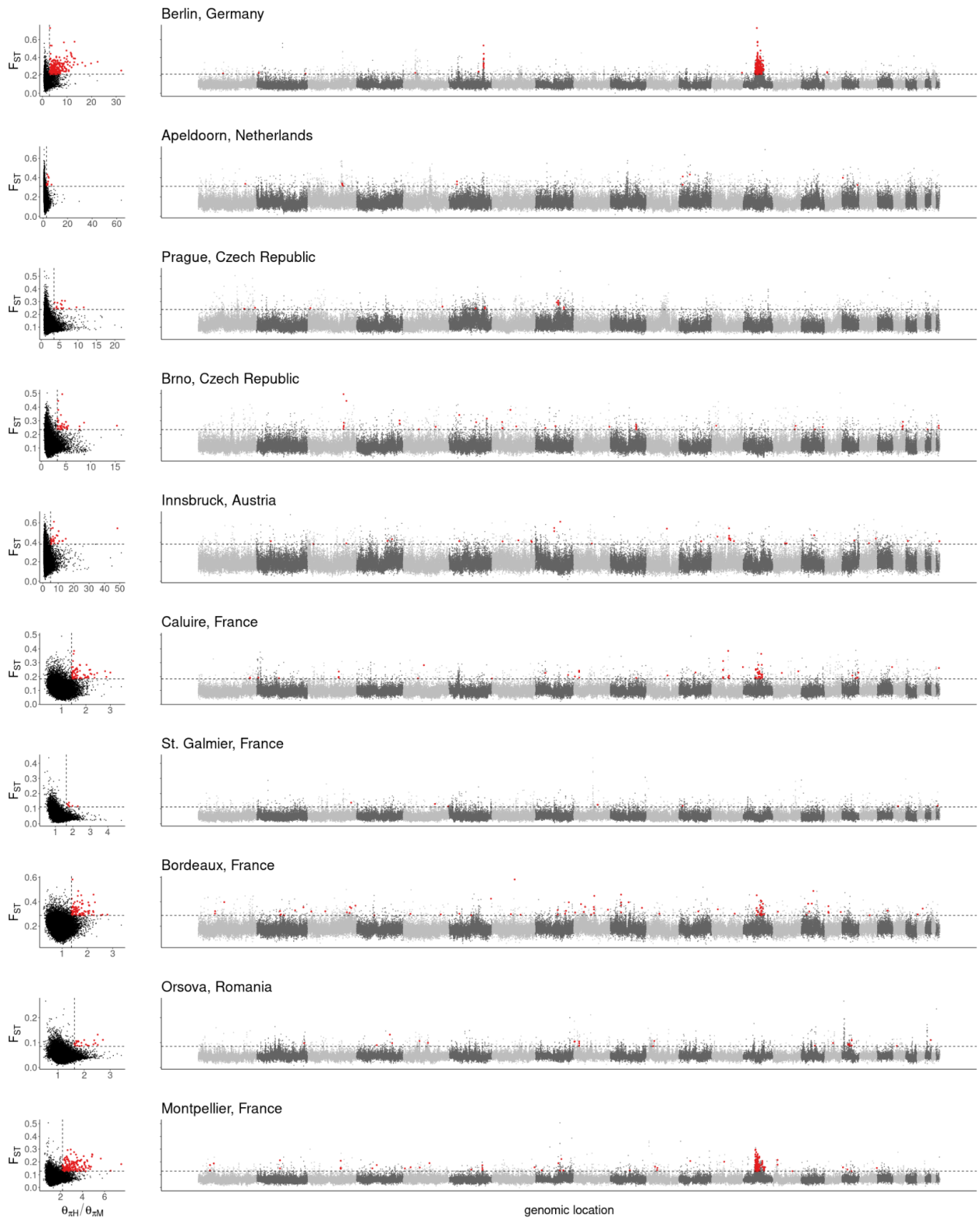

Figure S5. Distributions of F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ between historic and modern samples from European populations, and F_{ST} against genomic location. Red points indicate putative selective sweep windows, which are in top one percent of per-window F_{ST} and $\theta_{\pi H}/\theta_{\pi M}$ (dashed lines).

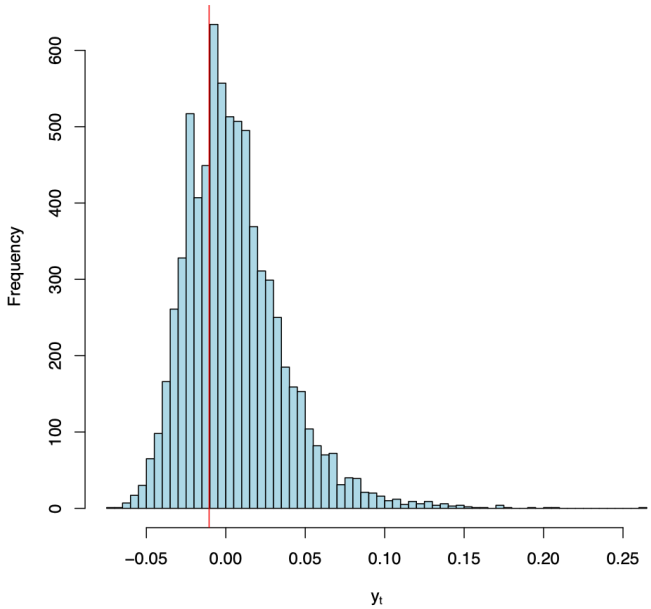

Figure S6. Distributions of y_t values (calculated according to equation 1) for allele frequency shifts in 10,000 putatively neutral SNPs (histogram) between historic and modern populations in Quebec. Red line indicates the y_t value for scaffold 2 position 56,025,662, a nonsynonymous SNP in *ELF3*.

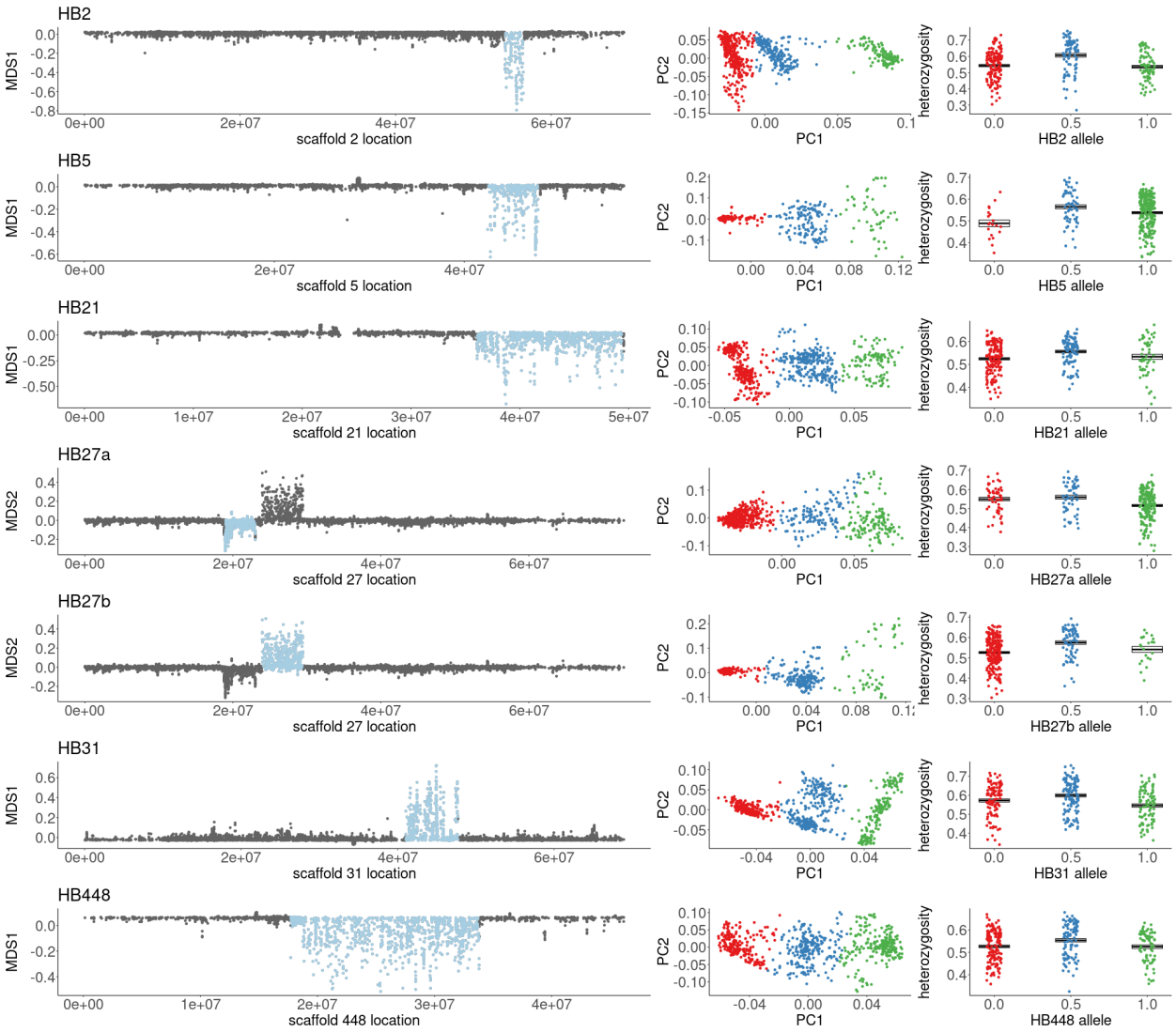

Figure S7. Haploblocks display extreme, divergent, local population structure (pale blue regions; left panes). Haploblock genotypes were assigned by kmeans clustering (colours; middle panes) using the first two principal components of genetic variation across haploblock regions. Heterozygous haploblock genotypes show elevated mean per-site heterozygosity (right panes; boxes denote mean and SEM).

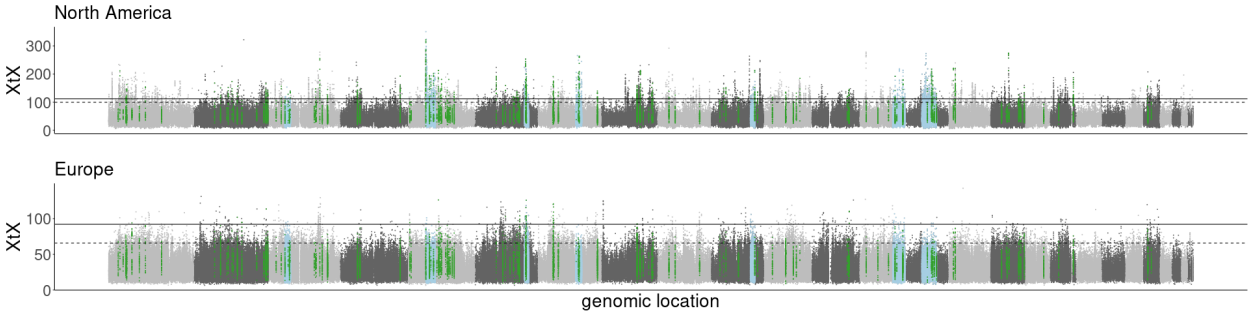

Figure S8. Genome-wide XtX scans between sampling locations within each range separately. Solid lines indicate Bonferroni-corrected significance derived from XtX *p*-values; dashed lines indicate the top 1% of genome-wide XtX values. Dark green highlights represent 10kb WZA windows with a top 5% WZA score in both ranges for XtX and at least one environmental variable. Pale blue highlights indicate haploblocks that overlap parallel WZA windows.

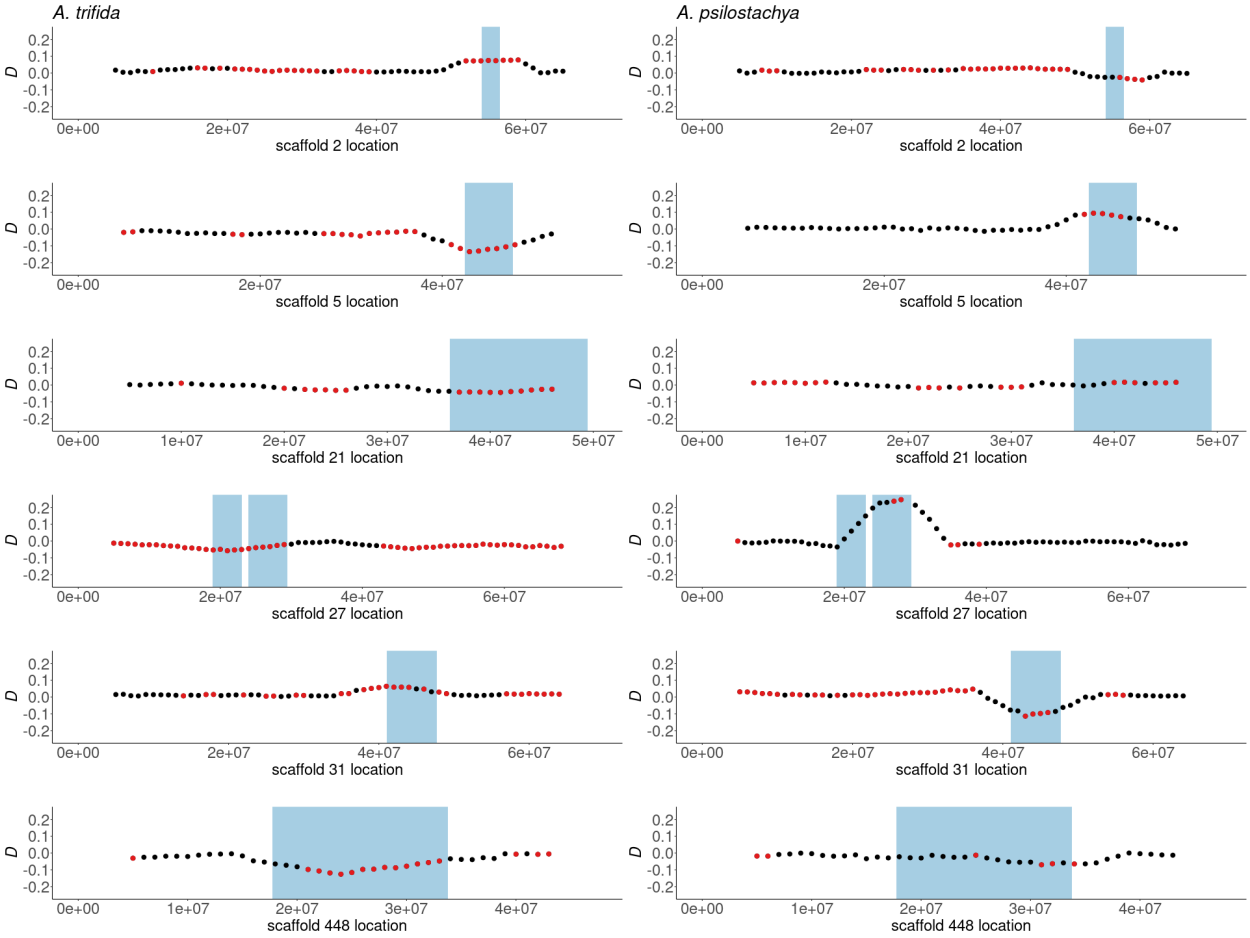

Figure S9. Patterson's *D* from ABBA-BABA tests of introgression using *Ambrosia trifida* and *Ambrosia psilostachya*, in 10Mb windows, for each scaffold harboring a haploblock. Haploblock regions are shown in pale blue; red points indicate windows with statistically significant *p*-values ($p < 0.05$).

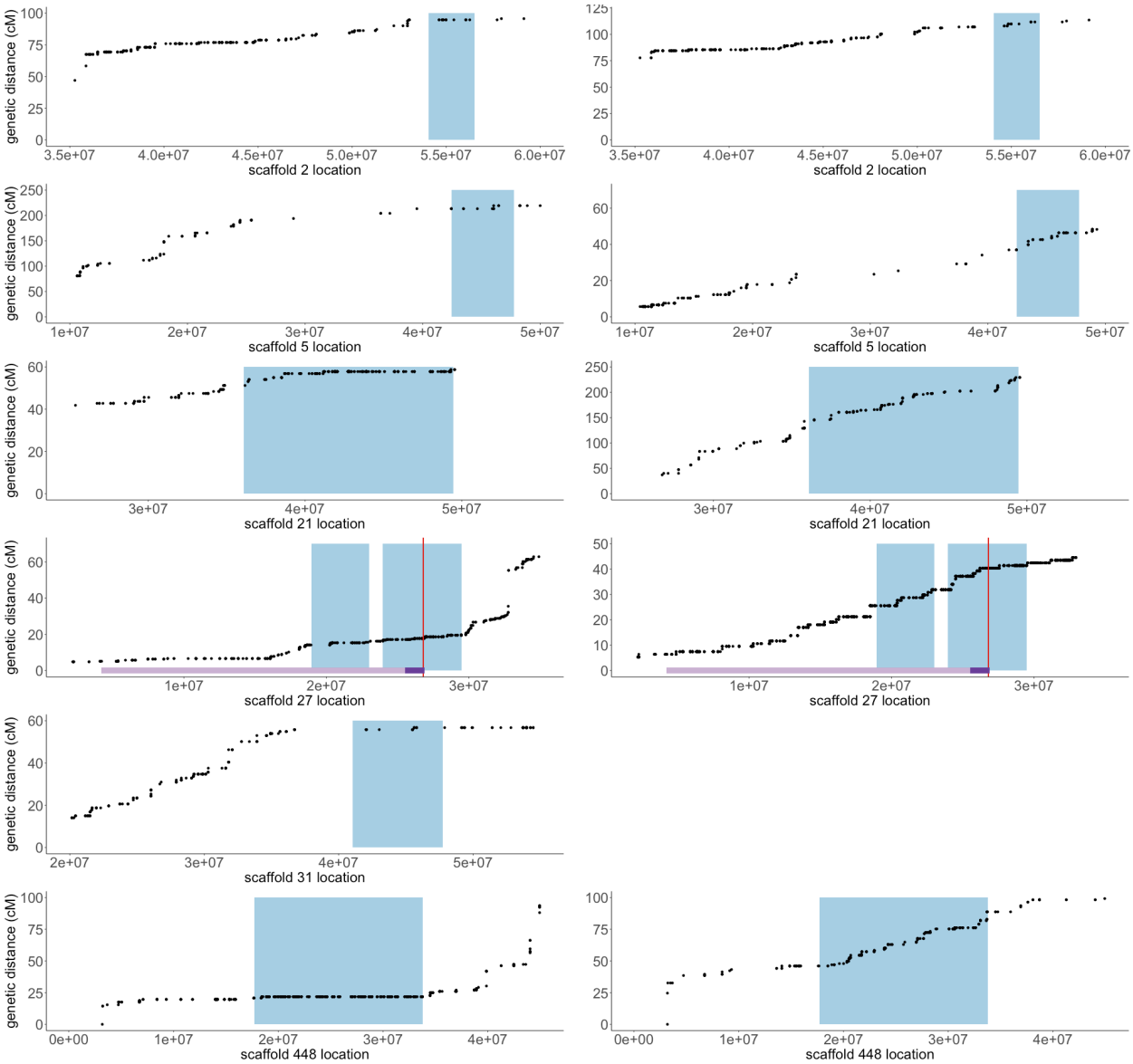

Figure S10. Evidence of genotype-specific reductions in recombination in haploblock regions. Genetic distance (cM) against physical distance (bp) along a portion of each scaffold is shown. Haploblock regions are shown in pale blue. Example maps are displayed showing both low recombination rates (left) 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 (*30*) 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.

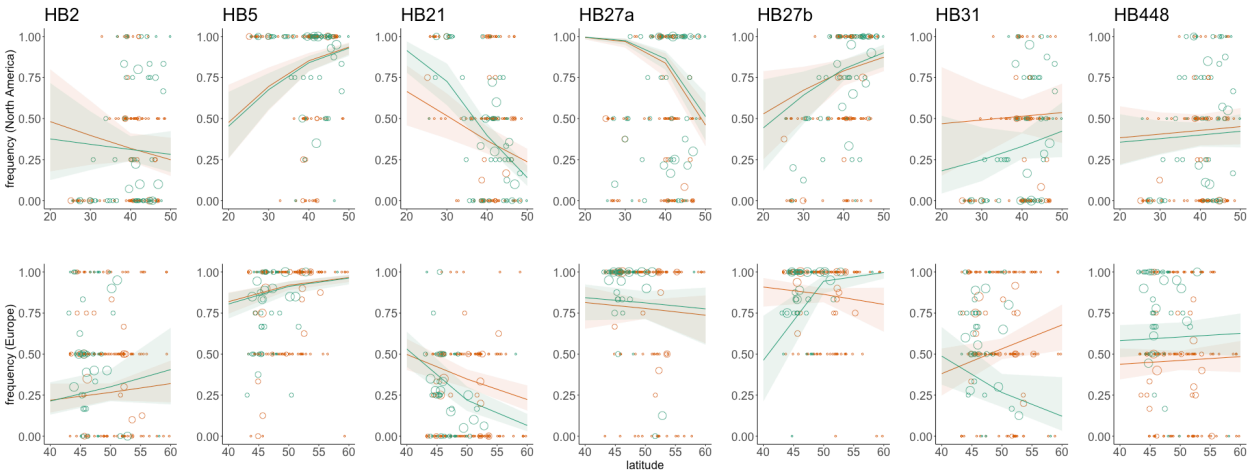

Figure S11. Frequency of haploblocks against sampling location latitude for historic (brown) and modern (green) *A. artemisiifolia* populations (circles; scaled relative to population sample size). Lines represent reduced generalized linear models with 95% CI ribbons (see tables S9-12 for model details).

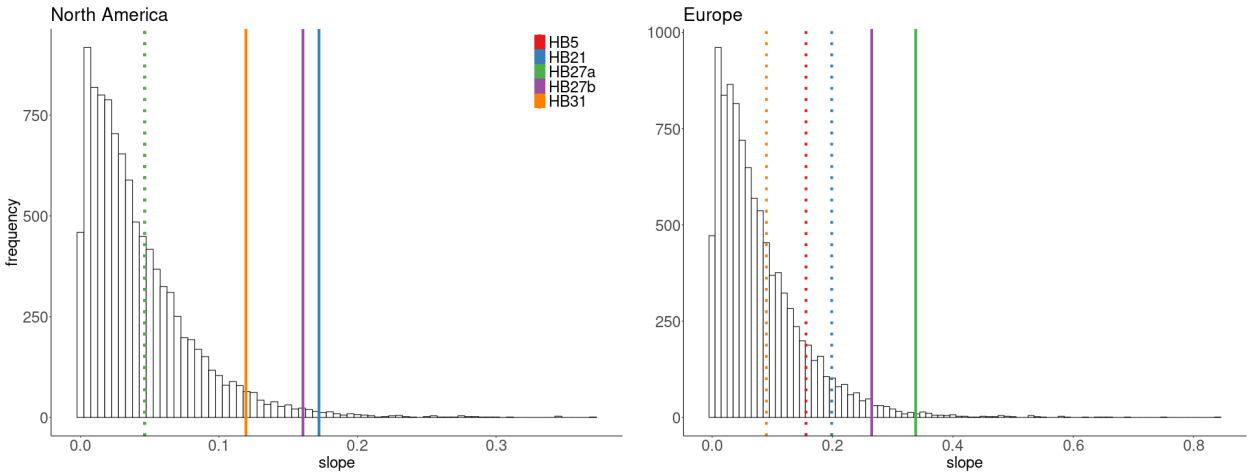

Figure S12. The distribution of slope estimates from generalized linear models of population allele counts against latitude for 10,000 randomly selected SNPs in each range. The vertical lines show the slope estimates for haploblocks with statistical associations with latitude in one range (table S9). Solid lines represent estimates in the 5% tail of each distribution while dotted lines fall below that cut-off.

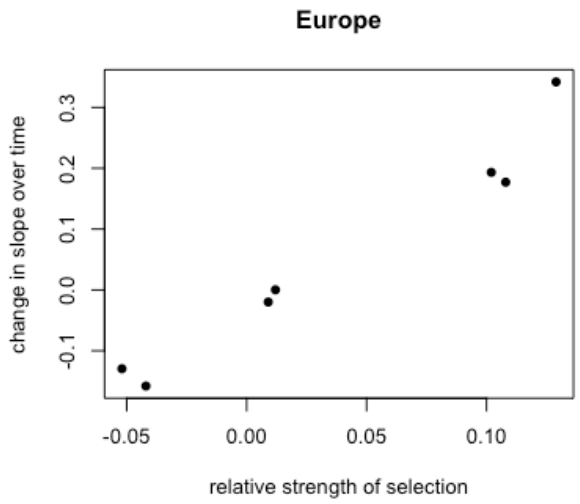

Figure S13. The change in the slope of the relationship between latitude and haplotype frequency (see table S12) between historic and modern samples compared to the estimate of selection along the latitudinal cline for the seven haploblocks (estimated from modern data in each range). A strong relationship was detected in the invasive European, but not the native North American range.

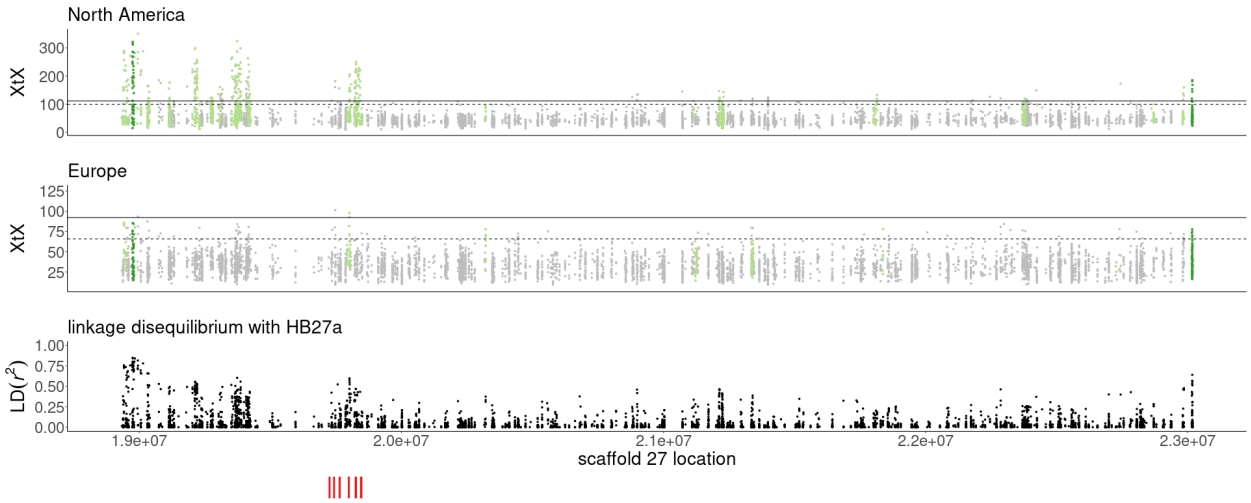

Figure S14. HB27a region (ScBFxKa_27:18930440-23019032). XtX and XtX-EAA outlier windows zoomed from fig. 1C, and linkage disequilibrium (r^2) between SNPs in the region and HB27a haploblock genotype. A cluster of pectate lyase genes, consisting of the top BLAST hit for *Amba1* and closely-related paralogues, are indicated in red below.

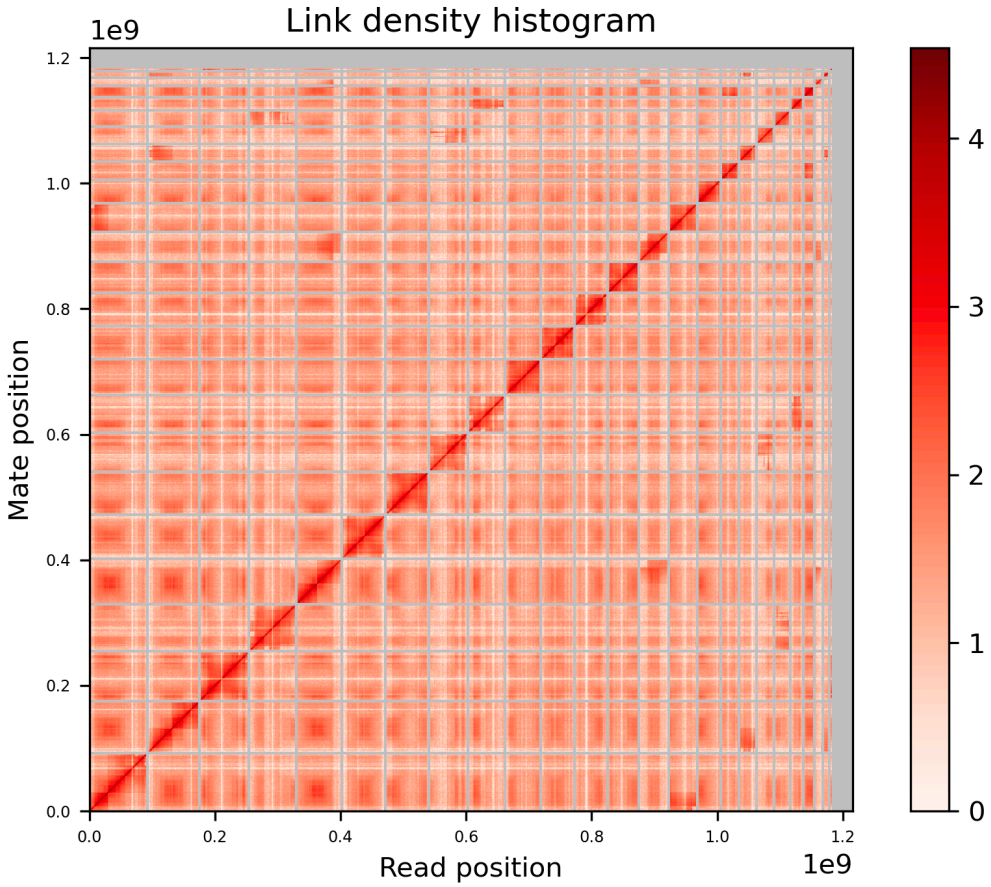

Figure S15. A link density histogram, identified by proximity ligation sequencing. The x and y axes show mapping positions of the first and second read in read pairs.

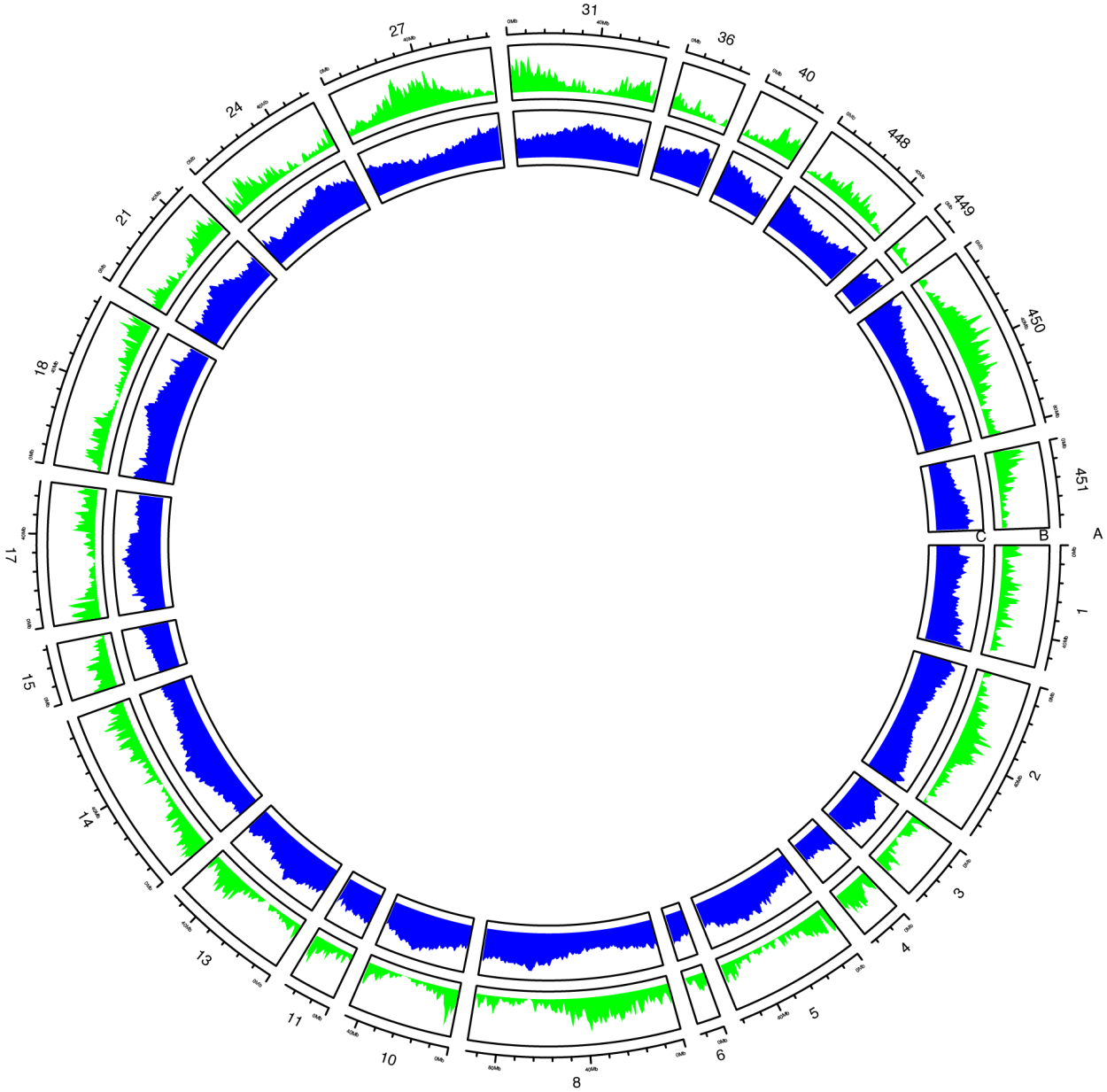

Figure S16. The *Ambrosia artemisiifolia* genomic landscape location across the 24 scaffolds greater than 10Mb in length (track **A**), and the distribution, in 1Mb windows of gene density (track **B**) and repeat density (track **C**).

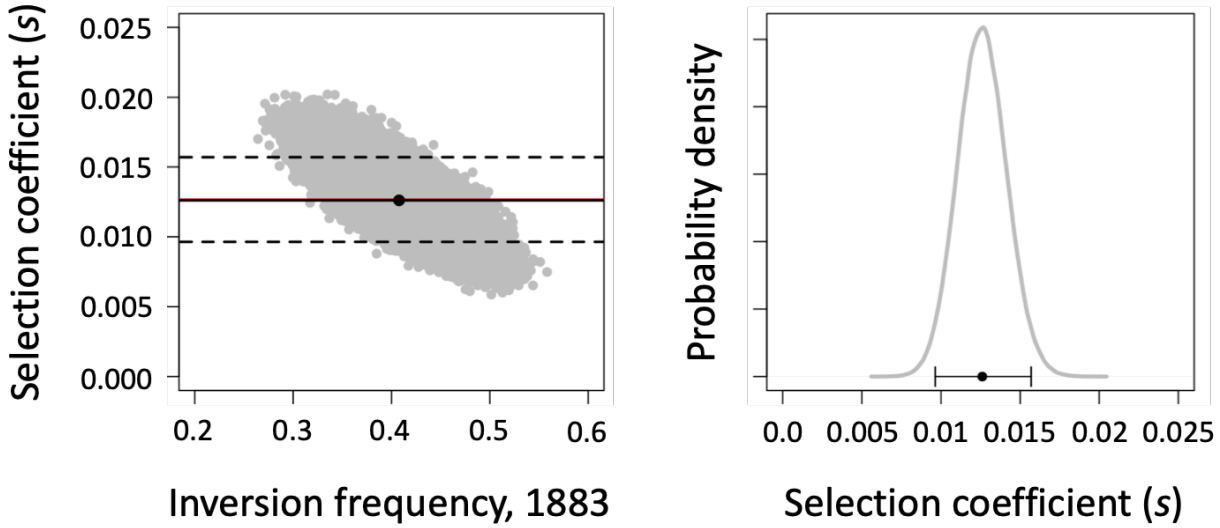

Figure S17. Selection coefficients consistent with observed frequency changes of the HB448 haplotype, under the positive selection model. The left-hand panel compares the distribution of *s* values estimated using approaches based on eqs. (S1) and (S2), as described in the surrounding text. The red line shows the mean estimate using eq. (S1) and the black line is the mean estimate using eq. (S2), with the dashed lines showing the 95% confidence interval of estimates based on the second approach. The x-axis represents the simulated values of the initial haplotype frequency, consistent with the empirical estimate of the initial frequency and uncertainty in the estimate. Gray circles show the relation between the simulated initial frequencies and the estimated selection coefficients. The black circle at the center corresponds to the empirical estimate of the initial frequency. The right-hand panel shows the distribution of estimated *s* values (gray) and its 95% CI (in black: parallel to the *x*-axis), which is a re-visualization of the data in the left-hand panel.

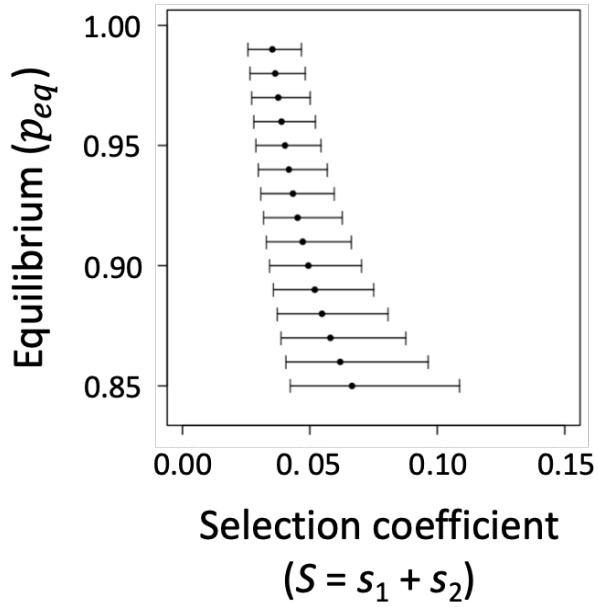

Figure S18. Selection coefficients in the balancing selection model that are consistent with observed changes in HB448 inversion frequencies in Europe. Equilibrium frequencies near unity (p_{eq} = 0.99) require modestly strong selection (average *S* = 0.035; 95% CI = [0.026, 0.047]) to explain the observed frequency changes. Lower equilibrium states (p_{eq} < 0.99) require stronger selection to explain the observed frequency changes in HB448.

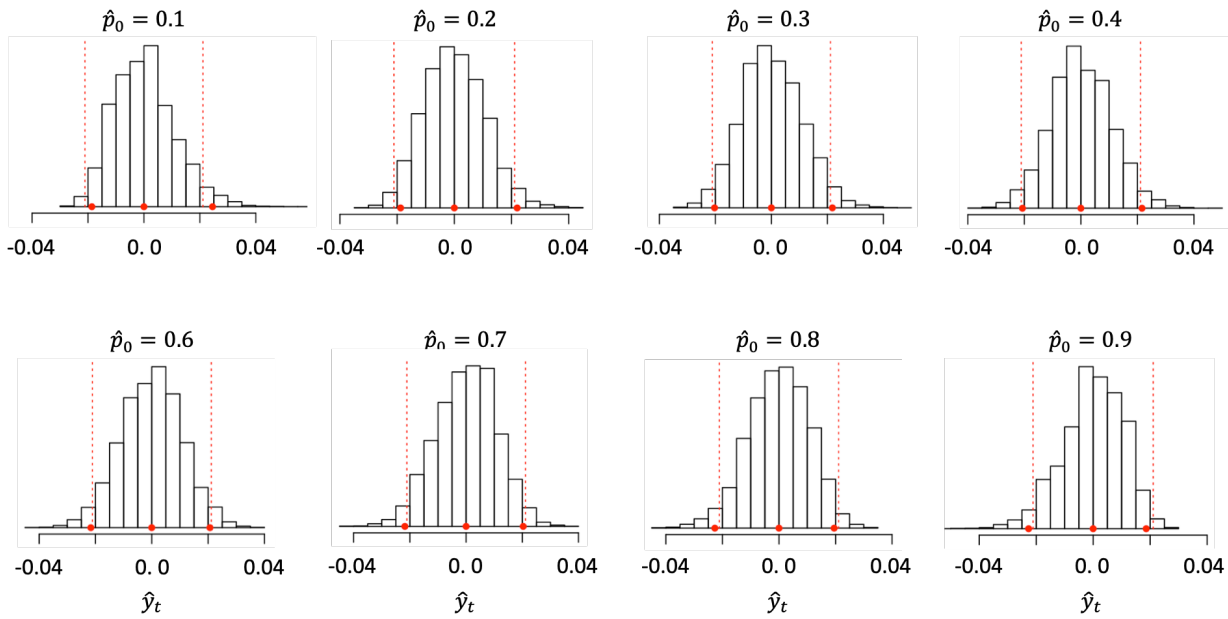

Figure S19. Simulated distribution of the scaled metric of divergence, $\hat{y}_t$, given different estimated values of the initial (historic) allele frequency. Each distribution is based on 10^4 independent and neutrally evolving SNPs. The simulations use n_0 = 182, n_t = 312, and t = 131, with a moderate 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, e.g. (*70 pp. 580-581*). 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.* (*71*) 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:

$$s = \exp\left[\frac{1}{t} log\left(\frac{p_t(1-p_0)}{p_0(1-p_t)}\right)\right] - 1 \qquad \text{(S1)}$$

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.

Selection can also be inferred by linear regression of the estimated inversion frequencies on the corresponding time points from which the estimates are derived. The logarithm of the ratio of inversion to non-inversion haplotype frequencies is a linear function of time, *i.e.*:

$$\log(x_t) = \log(x_0) + t\log(1+s) \qquad \text{(S2)}$$

in which $\log(x_0)$ represents the intercept and $\log(1+s)$ the slope in linear regression model.

We used both approaches 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 three time points: (1) 0.401 in the first historical sample with mean sampling date of 1883 (*n* = 91 diploid individuals; CI = [0.306, 0.504]), (2) 0.564 in the second historical sample with mean sampling date of 1930 (*n* = 102; CI = [0.467, 0.656]), and (3) 0.776 in a contemporary sample from 2014 (*n* = 156; CI = [0.704, 0.834]). 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 2*n* corresponding to the estimated frequency and sample size for the relevant time point (the factor of 2 in 2*n* accounts for diploidy).

We first used eq. (S1) to calculate 10^6 values of *s* that were compatible with simulated inversion frequencies for time points (1) and (3), and used these to infer the mean and 95% confidence interval for *s*, given our data. We also carried out 10^6 linear regressions of simulated inversion frequencies on the three time points (*i.e.*, *t* = 0, *t* = 1930 – 1883, and *t* = 2014 - 1883) to calculate the slope and intercept from the model in eq. (S2), and these were then used to infer the distribution of *s* (mean and 95% CI) consistent with the data. Both approaches yielded roughly the same results (fig. S17), with estimates of *s* = 0.0127 (CI = [0.0096, 0.0158]) based on the first approach and *s* = 0.0126 (CI = [0.0096, 0.0157]) based on the second. These results correspond to an estimated fitness difference between homozygotes with standard and inversion haplotypes of ~0.0255 and ~0.0254, respectively.

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 (*e.g.* *72*, *73*).

Following standard theory (*e.g.*, (*74*) 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:

$$S = \frac{p_{eq}\log\left(\frac{1-p_t}{1-p_0}\right) + \left(1-p_{eq}\right)\log\left(\frac{p_t}{p_0}\right) - \log\left(\frac{p_{eq}-p_t}{p_{eq}-p_0}\right)}{p_{eq}\left(1-p_{eq}\right)\left(t + \log\left(\frac{1-p_t}{1-p_0}\right) + \log\left(\frac{p_t}{p_0}\right) - \log\left(\frac{p_{eq}-p_t}{p_{eq}-p_0}\right)\right)} \quad \text{(S3)}$$

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. (S3) 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 exceeding 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. S18). 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; *75 p. 204*).

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 (*34*, *76 pp. 240-244*). 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"; see *77*). At equilibrium between selection and migration, the maximum cline slope will be:

$$\left(\frac{dp(x)}{dx}\right) = \sqrt{\frac{s}{3\sigma^2}} \tag{S4}$$

Following (*76 p. 244*), the equilibrium general solution for the cline is:

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

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

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:

$$\lim_{x \to 0} \frac{d \log\left(\frac{p(x)}{1 - p(x)}\right)}{dx} = 4\sqrt{\frac{s}{3\sigma^2}}$$

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

$$\lim_{x \to 0} \frac{d \log\left(\frac{p(x)}{1 - p(x)}\right)}{dx} = \sqrt{\frac{2s}{\sigma^2}}$$

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 $\sqrt{s}/\sigma$ by multiplying the values for β_1 by $\sqrt{3}/4$.

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:

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

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:

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

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

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:

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

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

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:

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

$$\begin{aligned}\mathrm{var}[\hat{p}_t|\hat{p}_0] &= \mathrm{E}\{\mathrm{var}[\hat{p}_t|p_0]|\hat{p}_0\} + \mathrm{var}\{\mathrm{E}[\hat{p}_t|p_0]|\hat{p}_0\} \\ &= \mathrm{E}\left\{\frac{p_0(1-p_0)}{n_t} + \frac{p_0(1-p_0)}{2N_e}t\left(1-\frac{1}{n_t}\right)\middle| \hat{p}_0\right\} + \mathrm{var}\{p_0|\hat{p}_0\} \\ &= \hat{p}_0(1-\hat{p}_0)\left[\frac{1}{n_t} + \frac{1}{2N_e}t\left(1-\frac{1}{n_t}\right)\right]\left(1-\frac{1}{n_0}\right) + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}\end{aligned}$$

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

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

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

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: $\mathrm{var}[\hat{y}_t] = \mathrm{var}[\hat{y}_t|\hat{p}_0]$.

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

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

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 $\mathrm{var}[\hat{y}_t|\hat{p}_0]$.

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; *78*). 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. S19 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 $\mathrm{var}[\hat{y}_t]$ (see above).