Large haploblocks underlie rapid adaptation in an invasive weed

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

These authors contributed equally: Paul Battlay, Jonathan Wilson
These authors jointly supervised this work: Kathryn A. Hodgins, Michael D. Martin

\textsuperscript{1} School of Biological Sciences, Monash University, Melbourne, Victoria, Australia
\textsuperscript{2} Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
\textsuperscript{3} Center for Evolutionary Hologenomics, GLOBE Institute, University of Copenhagen, Denmark
\textsuperscript{4} Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), AIMST University, 08100 Kedah, Malaysia
\textsuperscript{5} Department of Crop Production Ecology, Uppsala Ecology Center, Swedish University of Agricultural Sciences, Uppsala, Sweden.
\textsuperscript{6} Department of Grapevine Breeding, Hochschule Geisenheim University, Geisenheim, Germany.
\textsuperscript{7} Cell, Molecular Biology and Genomics Group, Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
\textsuperscript{8} Department of Environmental and Plant Protection, University of Novi Sad, Novi Sad, Serbia
\textsuperscript{9} Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, Canada
ABSTRACT

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

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

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

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

*Ambrosia artemisiifolia* populations are characterized by strong local adaptation and high gene flow between populations\(^{10,11}\). In Europe, invasive populations have been established through multiple genetically diverse introductions from North America over the last ~160 years\(^{11–13}\).

Remarkably, latitudinal clines observed for multiple traits in the native range, including flowering time and size, have re-evolved in Europe and Australia\(^{14}\), suggesting rapid local adaptation following invasion. Moreover, this trait-level parallelism is echoed in signals of parallelism at the genetic level\(^{10}\).

As biological invasions continue to increase worldwide\(^{15}\) and the effects of anthropogenic climate change intensify, understanding the genetic architecture of adaptation to sudden environmental shifts—a classical question in evolutionary research—becomes ever more important. While long-standing theory suggests that evolution in response to incremental environmental change should proceed through mutations of infinitesimally small\(^{16}\) or moderate\(^{17}\) effect, large-effect mutations are predicted to be useful in bridging extreme, sudden environmental shifts\(^{18}\). Moreover, alleles of large effect will, in cases of local adaptation, be better able to persist in the face of the swamping effects of gene flow\(^{19}\). Large-effect mutations are also more likely to produce patterns of evolutionary repeatability, or genetic parallelism, between species’ ranges\(^{20}\), particularly if adaptive responses make use of standing genetic variation (as would be expected during a bout of rapid adaptation), rather than *de novo* mutations\(^{21}\). These features of large-effect mutations may, however, be achieved by groups of mutations in tight genetic linkage\(^{19}\), including mutations captured by chromosomal inversions\(^{22}\).

There is substantial empirical evidence for the involvement of inversions in local adaptation\(^{23}\). For example, *Drosophila melanogaster*’s *In(3R)Payne* inversion shows parallel environmental associations across multiple continents\(^{24}\), and several plant inversions have been identified as contributing to local adaptation and ecotype divergence\(^{25,26}\). Theory also predicts that inversions can drive range expansions\(^{27}\), though their actual contributions to biological invasions are not well-understood.

Here we develop a new, high quality and near chromosome-level reference assembly, and examine genome-wide variation in over 600 modern and historic *A. artemisiifolia* samples from throughout North America and Europe\(^{11}\). Using this data of unparalleled spatial and temporal
resolution, we first identified regions of the genome experiencing climate-mediated selection in the native North American and introduced European ranges leveraging landscape genomic approaches and genome-wide association studies of adapting traits such as flowering time. Second, motivated by evidence that European and North American populations show similar trait clines with respect to climate, we examined the extent of between-range parallelism at the genetic level. Although adapting traits such as flowering time and size are polygenic, we expected to see substantial levels of parallelism if large and moderate effect standing variants were contributing to adaptive divergence. Third, we determined if these same regions show temporal signatures of selection in Europe, which would be expected if some invading populations were initially maladapted to their local climates. We coupled this temporal genomic analysis with a temporal analysis of phenological trait changes in European herbarium samples to further support our findings of genomic signatures of selection on flower time genes. Finally, we identified haplotype blocks with multiple features consistent with inversions, in genomic regions enriched for signatures of parallel adaptation. To determine if these putative structural variants were contributing to rapid local adaptation in Europe, we assessed spatial and temporal changes in their frequency as well as their associations with adapting traits.
RESULTS

Reference genome assembly

We assembled a new, near chromosome-level *Ambrosia artemisiifolia* reference genome from a heterozygous, diploid individual collected from Novi Sad, Serbia. After purging uncollapsed heterozygosity and scaffolding with HiRise our final assembly consisted of 395 scaffolds, with an \( L_{90} \) of 20 (A. artemisiifolia has 18 chromosomes and a genome size of 1.2Gb (flow cytometry estimates of genome size range from 1.13-1.16Gb)). A total of 253 complete Benchmarking Universal Single-Copy Orthologs (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 S1) were identified. The high numbers of duplicated BUSCO genes likely reflect the whole genome duplications experienced in the Asteraceae, including a recent one shared by *Helianthus* (sunflower) species at the base of the tribe. This species also retained a large number of duplicated BUSCO genes. A large fraction of the genome consisted of repetitive sequence (67%; table S2). Retroelements were the largest class (39.5%), with LTRs, particularly Gypsy (7.87 %) and Copia (18.98 %), the most prevalent retroelements. MAKER identified 29,849 gene models with strong protein or transcript support, with average coding lengths of 1.09kb, and 7.18 exons per gene (table S3; fig S1).

Genome-wide association studies

Genome-wide association studies (GWAS) using 121 modern samples across North American (\( n = 43 \)) and European (\( n = 78 \); table S4) ranges identified significant associations with 13 of 30 phenotypes, many of which are putatively adaptive, previously measured by van Boheemen, Atwater and Hodgins (fig. S2; table S5). 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*), an "evolutionary hotspot" for parallel flowering time adaptation in *A. thaliana*, barley and rice (table S5; 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). While the latitudes of these populations are greater in Europe than North America, the climatic conditions are similar (fig. S3), indicative of local climate adaptation in parallel between ranges.

Environmental-allele associations

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)\textsuperscript{39}, and correlated with environmental variables (latitude, longitude and 19 WorldClim temperature and precipitation variables; table S6)\textsuperscript{40}. Statistics were analyzed in 10kb windows using the Weighted-Z Analysis (WZA)\textsuperscript{41}. 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 S6; fig. 1C). This suggests that North America exhibits greater population differentiation, and a stronger relationship between population differentiation and the environment than Europe, which 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.

Previous studies in \textit{A. artemisiifolia} have identified signatures of repeatability between native and invasive ranges at phenotypic and genetic levels\textsuperscript{10,14}. 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 \( p = 4.39 \times 10^{-63} \); fig. 1C), with 19\% of climate adaptation candidates in Europe also candidates in North America. To account for the possibility that the number of parallel windows is inflated by extended linkage disequilibrium between windows (and hence represents 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 \( p = 1.08 \times 10^{-63} \)). Consequently, many of the same regions of the genome are involved in climate adaptation in both ranges.

In addition to \textit{ELF3}, parallel XtX-EAA outlier windows included four windows within two previously identified flowering-onset QTL\textsuperscript{42}. 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 S7). 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 S8). Some cytochrome P450 genes are involved in detoxification of xenobiotic compounds and the synthesis of defense compounds, while others
play key developmental roles\textsuperscript{43}, including contributing to the biosynthesis of gibberellin, a hormone that regulates a range of developmental events, including flowering\textsuperscript{44}. 
Figure 1. Signatures of climate adaptation in *Ambrosia artemisiifolia*. A. GWAS results (-log₁₀ 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 phenotypic analysis

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 985 digitized herbarium images (fig. S4) identified a significant shift in the probability of flowering and fruiting over time in Europe, but this change depended on the latitude or the day of the year the sample was collected (fig. S5; table S9). 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. For this trait, collection day also significantly interacted with collection year and more recently collected plants were more likely to be flowering later in the year and less likely to be flowering earlier in the year. 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. S5; table S10). This substantial spatio-temporal change in phenology is consistent with experimental common gardens that show that earlier flowering has evolved in northern populations and later flowering in southern populations following the invasion of Europe\textsuperscript{14}.

Further, this shift in both flowering and fruit set over time supports the hypothesis that an initial mismatch between the local environment and the genotypes present impacted the reproductive output of *A. artemisiifolia* during the early stages of colonization, particularly in northern Europe.

Temporal genomic signatures of selective sweeps

Genome resequencing of *A. artemisiifolia* herbarium samples\textsuperscript{11} allowed comparisons between historic and modern populations. We grouped historic samples based on their age and proximity to a modern population sample, resulting in five North American and seven European historic-modern population pairs (table S11), 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 (494 unique windows) than in North America (113 unique windows; fig. S6 c.f. S7; table S11), 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 two French populations in a ~16Mb region on scaffold 448 (fig. 2A; fig. S7), which accounts for 312 (63\%) of the European sweep windows. In Berlin, sweep windows were also observed containing and surrounding the flowering onset GWAS peak that includes *ELF3*. In comparisons of spatial and
temporal signatures of selection, two and nine sweep windows were also XtX-EAA outliers in North America and Europe respectively. All such windows in Europe were in the scaffold 448 or ELF3 regions. To further investigate the temporal shift associated with ELF3, we focused on the nonsynonymous ELF3 variant across all samples within a 200km radius of Berlin (15 historic and eleven modern samples). The frequency of the variant increased from 4.9% to 73.9% between historic and modern samples, an allele frequency shift greater than 10,000 putatively neutral loci sampled from the same geographic region (fig. 2C); similar shifts were not observed in North American samples within a 200km radius of Quebec City (ten historic and eleven modern), where the ELF3 allele is at comparably high frequencies (fig. S8).
Figure 2. Temporal signatures of selective sweeps in Europe. A. Distributions of $F_{ST}$ between historic and modern samples and the ratio of historic to modern nucleotide diversity ($\theta_{mH}/\theta_{mM}$) 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 $\theta_{mH}/\theta_{mM}$ (dashed lines). B. Strong evidence for a selective sweep on scaffold 448 in European populations corresponds with local divergent population structure (MDS1), indicating the presence of a haploblock (putative chromosomal inversion) in this region. C. Distributions of values of a standardized measure of allele frequency change, $y_t$ (calculated according to equation 1; histograms) for shifts between historic and modern populations across putatively neutral SNPs. $y_t$ for scaffold 2 position 56,025,662, a nonsynonymous SNP in ELF3 (red line), against a distribution of $y_t$ values from 10,000 putatively neutral SNPs from samples within 200 km of Berlin (left) and $y_t$ calculated from the haplotype frequency of HB448 (red line), against a distribution of $y_t$ values from 7,394 putatively neutral SNPs from across the European range (right).
**Haploblock identification**

Chromosomal inversions have previously been identified as driving local adaptation of ecotypes of *Helianthus* species\(^2\). We used a similar approach to identify genomic signatures of putative inversions (haploblocks) contributing to local adaptation in *A. artemisiifolia*. Briefly, we identified genomic regions in which population structure was divergent and fell into three clusters, putatively representing the heterozygous and two homozygous genotypic classes of an inversion. Further, we looked for pronounced shifts in population structure (indicating inversion breakpoints), elevated local heterozygosity in the heterozygous cluster, and increased linkage disequilibrium across the region (fig. S9). Finally, we examined mapping populations of *A. artemisiifolia*\(^4\) for evidence of map-specific reductions in recombination across haploblock regions (i.e., suppressed recombination in haploblock regions in some maps but not others; fig. S10). This would be the pattern expected when recombination is suppressed by inversions in heterozygotes but not homozygotes, as opposed to the haploblocks being caused by global reductions in recombination in those regions. All haploblocks showed evidence of suppressed recombination in some maps but not others, with the exception of HB31 which showed suppressed recombination in all maps.

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

**Haploblock frequency changes through space and time**

To identify changes in haploblock frequency over time and space, which would be consistent with selection on these putative inversions, we first estimated haploblock genotypes for all historic and modern samples. Within haploblock boundaries identified using modern sample SNP data in Lostruct\(^5\), we performed local PCAs with both historic and modern samples (table S4) and identified genotypes by k-means clustering (fig. S9). For modern samples, we used generalized linear models to estimate the slopes of the haploblock frequencies as a function of latitude within each range. For those haploblocks that were significantly associated with latitude,
we compared these estimates with the genome-wide distribution of slopes for North America and Europe, based on 10,000 unlinked SNPs that were randomly selected from outside haploblocks and genes. The estimated slopes for haploblocks HB27a, HB27b and HB31 fell into the 5% tail of the distribution for at least one of the ranges (fig. S12). However, this approach did not examine temporal changes nor the combined signatures of selection over space and time. To do so, we ran generalized linear models comparing haplotype frequency with latitude, time (date of specimen in years) and range (North America vs. Europe; fig. 3A; fig. S13; table S12-S17). All of the haploblocks showed significant changes over time, either in their average frequency in one or both ranges, or in their relationship with latitude within each range, a pattern that is consistent with local selection on the haploblocks. All seven haploblocks showed significant associations with time or with latitude in at least one range, indicative of climate adaptation. These patterns were robust to time being coded as discrete (historic vs. modern) or as continuous (by year; table S12). In HB5, HB27b and HB31, we identified significant three-way interactions between latitude, time and range (fig. 3A). The HB27b haplotype frequency was correlated with latitude in modern and historic samples from North America, as well as in modern European samples, consistent with climate-mediated selection. However, historic European populations did not display an association with latitude, which may reflect maladaptation during the initial stages of the European range expansion (fig. 3A; table S15, S17). A significant strengthening of the relationship between latitude and haplotype frequency was also observed in HB5.

To further investigate whether European populations showed evidence of recent local selection on the haploblocks, we tested whether estimates of selection inferred from contemporary spatial data were associated with temporal changes in haploblock frequencies between historical and contemporary European populations. We used spatial variation in contemporary haploblock frequencies to estimate the relative strength of local selection on these haploblocks (see supplementary text S2). We specifically compared estimates of the maximum slope of latitudinal clines for each putative inversion’s frequency 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 strength of local selection for a given inversion and $\sigma$ is the average dispersal distance of individuals in the range$^{46}$ (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. We found that estimates of the relative strengths of...
local selection across the haploblocks were correlated between the ranges, indicating parallel
patterns of local selection along the latitudinal gradient ($r = 0.80$, $p = 0.03$). We also found that
changes in the haploblock cline slopes between historic and modern time points within Europe
were significantly correlated with our estimates of the relative strength of spatially varying
selection of the haploblocks across the European range ($r = 0.86$, $p = 0.01$ fig. S14). Such a
pattern is consistent with a scenario in which historical European populations were not initially
locally adapted (haploblock frequencies were initially far from local optima) and where the
haploblocks subject to relatively strong local selection exhibited the greatest temporal changes
in local frequency over the ensuing century. The same pattern was not observed in the North
American native range, whose historic populations are likely to have been consistently closer to
the local optima across the timescale of our analysis ($r = 0.61$, $p = 0.15$). Cline slopes for
latitude were also shallower in historic relative to modern European populations ($t_{6.58} = -2.39$, $p
= 0.05$; mean absolute slope: historic EU = 0.05; modern EU = 0.15), but not so in North
America ($t_{11.9} = 0.24$, $p = 0.81$), consistent with initial maladaptation in Europe, followed by
adaptation to local climates.

The HB448 haplotype frequency dramatically increased in the European invasive populations
over time, consistent with strong range-wide selection in Europe (table S12). The haploblock
also showed a significant range effect, with a significantly higher frequency in modern Europe
compared to North America (historic and modern) and historic Europe (table S14). Latitudinal
clines in frequency were also observed although the clinal patterns were similar over time and
between ranges. Signatures of selective sweeps were observed in the HB448 region in multiple
European populations (fig. 2A; fig. S10). To further distinguish the effects of drift from those of
selection, we compared an empirical null distribution of allele frequency changes over time in
Europe using 2,296 SNPs matched for allele frequency (fig. 2C). Only a single locus in this null
distribution showed an allele frequency change greater than HB448. We then estimated the
strength of selection for the putative inversion under scenarios of positive and balancing
selection (see supplementary text S1). With positive selection for the putative inversion, the
observed changes in frequency are consistent with a 2.2% difference in fitness (95% CI =
{1.6%, 2.8%}) between individuals homozygous for the inversion relative to homozygotes for the
alternative haplotype. Scenarios of balancing selection require stronger selection to explain the
inversion frequency shift over time (see supplementary text S1). This estimate is smaller than
many empirical estimates of selection on individual loci in natural populations47. The greater
timespan of our samples facilitated detection of strong signals of selection for loci that would
otherwise be missed by the short time periods afforded by most temporal studies.

### Biological functions of haploblocks

Several analyses provide evidence for the biological function of these putatively adaptive
haploblocks. From genome annotations, we found that haploblocks were collectively enriched
for flowering-time pathway genes (Fisher’s exact test \( p = 0.034 \)), although individual
haploblocks were not significantly enriched (table S18). HB448 was enriched for the “recognition
of pollen” gene ontology term, with 28 (17%) of the genome’s 167 genes annotated with this
term falling within this haploblock (table S19). 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. 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.
S15). HB27b corresponds to a large-effect flowering time and height QTL. 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;
four haploblocks showed associations with traits related to flowering time (fig. 3B; table S20).
Figure 3. Haploblock distributions and trait associations. A. Logistic regression models with 95% CI ribbons (see table S12-17 for model details) of haploblock frequency (allele 1) against latitude for four haploblocks across five time bins ranging from most historic (purple) to most modern (green). B. Examples of significant associations between haploblock alleles and phenotypes (boxes denote mean and SEM). C. HB27b allele frequency in modern *A. artemisiifolia* populations.
DISCUSSION

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.

While invasive species are often envisaged to encounter novel selection pressures as they spread across alien landscapes, they must also readapt to similar environmental variation encountered in their native range, as haphazardly-introduced invaders are unlikely to be well-adapted to local conditions when the invasive populations initially expand across climatic gradients. Much of *Ambrosia artemisiifolia*’s European invasion lies within climatic extremes encountered across its native range (fig. S3). Despite this similarity in climatic variation, the patterns of parallel climate adaptation between native and invasive ranges are striking given the evolutionarily recent introduction of the species into Europe. As *A. artemisiifolia*’s invasion of Europe consisted of multiple introductions over a brief evolutionary time scale, these patterns are likely examples of ‘collateral evolution’⁴⁹, in which standing genetic variation in *A. artemisiifolia*’s native range has been co-opted for adaptation in and across the European invasive range. 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 certain paths to adaptation are favored, such as when beneficial variants are already present in the population as standing variants, evolution will repeatedly draw on the same subset of genes to reach the same adaptive endpoints.

From herbarium specimens that were sampled throughout the course of *A. artemisiifolia*’s invasion of Europe, we observed an abrupt change in flowering and fruiting over time. Leveraging whole-genome sequences of herbarium samples across North America and Europe, we were also able to scan populations for temporal genomic signatures of selective sweeps. Although some populations have experienced shifts in ancestry over time in Europe¹¹, peaks against the genome-wide background provide compelling evidence for rapid local adaptation in European populations, with the strongest genetic signals of rapid change over time corresponding to some of the strongest signatures of local adaptation in our spatial analyses, particularly windows in the region of the *ELF3* gene and HB448. Further, these regions show parallel signals of climate adaptation in North America and are associated with adapting traits...
such as flowering onset. These multiple lines of evidence provide strong support that climate
mediated selection on phenology was pivotal in shaping the adaptive genetic landscape of *A. arte-
" misiifolia* in Europe.

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\[^{22,27}\] , providing a genetic architecture
suited to local adaptation in the face of high gene flow\[^{11,19}\]. 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\[^{23}\]. 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 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
phenotypic 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 that are key factors in the timing, length and severity of the local pollen season
(e.g. days to flowering onset, days to the end of pollen production, and biomass). Consequently,
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\[^{50}\].
METHODS

Genome assembly

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

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

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

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

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

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

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

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

Sample alignment, variant calling and filtering

FASTQ files from historic and modern \textit{A. artemisiifolia} samples from North America and Europe\textsuperscript{11} were aligned to our new reference genome using the Paleomix pipeline\textsuperscript{56}, which incorporates AdapterRemoval\textsuperscript{55}, BWA-MEM\textsuperscript{53}, Picard MarkDuplicates (https://broadinstitute.github.io/picard/) and GATK IndelRealigner\textsuperscript{57}. Mean depths of alignments ranged from 0.33X to 18.03X with a mean of 3.67X for historic samples, and 1.60X to 39.80X with a mean of 6.26X for modern samples (table S4). Variants were called in the higher-depth modern samples using GATK UnifiedGenotyper\textsuperscript{58}. GATK VariantFiltration\textsuperscript{57} and VcfTools\textsuperscript{59} 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\textsuperscript{60}.

Genome annotation

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.

Prior to annotation of the genome, repetitive elements were identified using RepeatModeler2. ProExcluder was then run to remove any protein coding genes from the repeat library.

RepeatMasker was used to mask the genome using the finalized repeat library (table S2). A large fraction of the genome consisted of repetitive sequence (67%; fig. S1). Retroelements were the largest class (39.5%), with LTRs, particularly Gypsy (7.87 %) and Copia (18.98 %), the most prevalent retroelements.

Genome annotation was performed using the genome annotation pipeline MAKER2 version 2.31.9 with ab initio and homology-based gene predictions. 3,819 unique UniProtKB/Swiss-Prot protein sequences 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 to train the ab initio gene predictor SNAP. 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\textsuperscript{11}.

A high confidence gene set of 29,849 gene models with strong protein or transcript support was identified. These gene models had average coding lengths of 1.09 kb and 7.18 exons per gene (table S3). Gene models were compared with \textit{Arabidopsis thaliana} annotations (TAIR10 representative gene model proteins\textsuperscript{67}) and the UniProtKB database using the \textit{blastp} command in BLAST\textsuperscript{+}. Using an E-value threshold of $1 \times 10^{-6}$, 26,688 (89.4\%) genes matched TAIR10 annotations and 23,019 (78.1\%) were matched UniProtKB (table S21). Gene ontology (GO) enrichment was assessed using GO terms from \textit{A. thaliana} TAIR 10\textsuperscript{67} BLAST results. To identify GO terms enriched among candidate lists, the R/topGO package\textsuperscript{69} 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 \textit{A. thaliana} FLOR-ID flowering time pathway genes\textsuperscript{70}. 538 predicted \textit{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\textsuperscript{71}.

\textbf{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\textsuperscript{72}, 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\textsuperscript{39}, with an $\Omega$ 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\textsuperscript{72}. To identify allele frequency variation associated with environmental variables within ranges, 19 bioclimatic variables were extracted for each sampling location from the WorldClim database\textsuperscript{40} using the R/raster package\textsuperscript{73}. 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\textsuperscript{74}. Genome-wide XtX and τ results
were analyzed in non-overlapping 10kb windows using the weighted-Z analysis (WZA)\(^{41}\), with the top 5% of windows designated outliers.

**Genome-wide association studies**

Imputed genotypes from modern samples were filtered in PLINK 1.9\(^{72}\). Non-SNP sites and sites with more than two alleles were removed. The 121 samples overlapping those phenotyped by van Boheemen, Atwater and Hodgins\(^{14}\) were retained (table S4), 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\(^{75}\), and incorporating an identity-by-state kinship matrix (generated in PLINK 1.9)\(^{72}\) 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\(^{14}\).

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. S4). 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)\(^{76}\) 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\textsuperscript{14} and we expected northern populations to evolve early flowering relative to the start of the growing season to match the shorter growing seasons in these areas. As a result, if local phenology has evolved to better match the local growing season we predicted a collection year by latitude interaction, as the relationship between latitude and the probability of flowering in wild collected accessions should change over time when controlling for the day of collection.

**Historic-modern genomic comparisons**

To identify targets of recent selection, we compared historic and modern samples from twelve locations (five locations from North America and seven from Europe; table S11). Historic samples were grouped based on age of sample and proximity to a modern population. Analyses were performed in ANGSD\textsuperscript{77} using genotype likelihoods. For each population location we calculated pairwise nucleotide diversity ($\theta_{\text{π}}$) for historic and modern populations separately, and $F_{\text{ST}}$ between historic and modern populations at each location. Statistics were calculated in non-overlapping 10kb windows, and windowed $\theta_{\text{π}}$ values were normalized by dividing by the number of sites in each window. At each location, windows with $\theta_{\text{π}}$ 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 $F_{\text{ST}}$ and $\theta_{\text{πH}}/\theta_{\text{πM}}$ 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 \textit{ELF3} non-synonymous SNP and the haploblock HB448, which are two candidate loci for recent selection in Europe. Both candidates showed evidence of local selection using spatial analysis of modern populations, as well as sweep signals in temporal comparisons of individual populations. These calculations were performed in geographic regions where this recent selection is believed to have occurred at both historic and contemporary timepoints. ANGSD (-\textit{minMapQ} 10 -\textit{minQ} 5 -\textit{GL} 2 -\textit{doMajorMinor} 1 -\textit{doMaf} 2 -\textit{doIBS} 1 -\textit{doCounts} 1 -\textit{doGlf} 2) was used to calculate the allele frequency of the early flowering \textit{ELF3} allele (ScBFxKa_2:56025662) in 15 historic and eleven modern samples from within 200km of Berlin, whilst the frequency of HB448 in Europe was ascertained using haploblock frequency estimates from across the European range (see below). To understand the
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:

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

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 1.972. 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 within 200km of Quebec City (ten historic and eleven modern) 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 Lostruct45. Using SNP data from 311 modern samples, we extracted the first ten multidimensional scaling (MDS) coordinates across each of the 30 imputed scaffolds (scaffolds with > 100 SNPs remaining after filtering) 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 X\(X\) 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
\(R^74\) to identify regions containing three distinct clusters representing heterozygotes and two
homzygotes. Additionally, we assessed heterozygosity from genotype data in each haploblock
region and in each modern sample, and measured linkage disequilibrium (the second highest \(r^2\)
value in 0.5Mb windows) across each scaffold bearing a haploblock for all modern samples and
for modern samples homozygous for the more common haploblock genotype using scripts from
Todesco et al.\(^26\).

**Haploblock frequency changes over time and space**

For seven candidate inversions, a local PCA of each region and kmeans clustering was then
repeated in PCAngsd\(^78\), so as to allow genotype estimation of these haploblocks in 305 historic
samples alongside the 311 modern samples. For this local PCA we used sites diagnostic of the
haplotypes (with \(F_{ST}\) values > 0.1 between the identified homozygote clusters in modern
samples, using vcftools v0.1.15 –weir-fst-pop (Auton and Marcketta, 2009) confined to the
haploblock coordinates) to reduce clustering ambiguity which was heightened by the low
coverage historic samples. We also conducted a PCA on 10,000 SNPs randomly-sampled from
the 311 modern genomes 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\(^72\). Following this, we used generalized linear models (glm \(R\)) to assess how
haplotype frequency (binomial response) changed over time and space. A count of each
haplotype at a geographic location and year 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 from the PCA of 10,000 randomly-
sampled SNPs was included as a covariate 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 \(R\))\(^76\) with type 3 tests. Significant differences among groups for means or
slopes were tested with the emmeans package using an FDR correction\(^79\). To determine if the
classification of samples into modern or historic timepoints influenced our results we ran a
second set of generalized linear models examining haplotype frequency as a function of
collection year, range (North America or Europe), latitude, and all interactions between these
three main effects as well as PC1, using the same approach as above. For interactions
involving two continuous variables (i.e., latitude and year) we tested if the slope estimates of
one variable were significant at specific values of the other using the package emmeans. This
allowed us to estimate when and where the haplotype frequencies were changing. The results
from both approaches (time as two categories or time as continuous) provided qualitatively
similar patterns.

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 (modern or historic) combination (group). 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 R79; table S13). PC1 was included as a covariate 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. We also expected that the magnitude of change in the slope over time
would be the greatest in haploblocks showing the largest estimates of selection in Europe (table
S15). 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 genes and haploblocks. We did this to
determine if our haploblocks showed stronger latitudinal patterns than the majority of SNPs, in
one or both ranges, 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 null SNP and each haploblock that was statistically
associated with latitude.
Recombination rates in haploblocks

The haploblocks show multiple genomic signatures of reduced recombination. To confirm this, we analyzed recombination rates in genetic maps. Further if the haploblocks were caused by global reductions in recombination rate (e.g., the region was found in an area with generally low recombination such as a centromere), all maps should show reduced recombination rates.

However, if inversions were the cause, recombination would only be suppressed in genotypes heterozygous for the inversion, while homozygous individuals would not show suppressed recombination. 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\(^42\). 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 Prapas et al.\(^42\). We developed sex-specific genetic maps (i.e., maps for the maternal and paternal parent) using Lep-MAP\(^38\) 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 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 QTL and the boundaries of the haploblocks were visualized and inspected for reduced recombination compared to the rest of the scaffold.
We thank Greg Owens & Michael Whitlock for discussions, and Samuel Yeaman and Sarah Otto for feedback on the manuscript. We are grateful to François Bretagnolle, Myriam Gaudeul, Heinz Mueller-Schaerer, Gerhard Karrer, and Bruno Chauvel for their assistance in obtaining many of the samples upon which this study is based, and Marie Brunier, Fátima Sánchez Barreiro, Yohann Roy, Luna Forcioli, Jacqueline Y. Lee for assistance during lab work. Some sequencing services were provided by the Norwegian Sequencing Centre, a national technology platform hosted by the University of Oslo. Some sequencing was performed by the NTNU Genomics Core Facility. Genome scaffolding was performed by Dovetail Genomics. Some analyses were performed on resources provided by Sigma2, MASSIVE M3 and ComputeCanada high performance computing platforms. We kindly thank the curators from the following herbaria for allowing us to destructively sample their precious collections: B, BR, BRNU, C, FI, G, GH, GOET, GZU, HBG, I, IASI, JE, L, LD, LY, MARS, MASS, MO, MPU, NEBC, NEU, NY, P, PH, PR, PRA, PRC, QFA, S, STU, TRH, UPS, US, W, WU.

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

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

**COMPETING INTERESTS**

The authors declare no competing interests

**DATA AND CODE AVAILABILITY**

Sequences used in reference genome assembly and annotation are available from NCBI under BioProject ID PRJNA819156 (to be released upon publication; reviewer link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA819156?reviewer=ppu225pmsn5tbfvf04s3tadkfj). Reference genome FASTA and annotation GFF files are available from FigShare (to be released upon publication). Individual sample resequencing data are available from ENA under BioProject IDs PRJEB48563, PRJNA339123 and PRJEB34825.
Figure S1. The Ambrosia artemisiifolia genomic landscape location across the 24 scaffolds greater than 10Mb in length (track A), and the distribution, in 1Mb windows of gene density (track B) and repeat density (track C).
Figure S2. Genome-wide association study results for phenotypes with significant SNPs. Solid lines indicate a Bonferroni-corrected significance threshold of 0.05.
Figure S3. Environmental comparison of North American and European A. artemisiifolia ranges. The left pane shows the first principle component of 19 bioclimatic variables against latitude for modern ragweed populations in North America (red) and Europe (blue). Crosses identify the populations with high ELF3 allele frequencies. The right pane shows a variable correlation plot for 19 bioclimatic variables.
Figure S4. An example herbarium specimen of *Ambrosia artemisiifolia* (left) and a detail (right) showing the mature male inflorescence (solid line) and seeds (dashed lines).
**Figure S5.** Interaction plots illustrating the results of generalized linear models examining the presence of mature male inflorescences (probability of flowers) or mature fruit (probability of fruit) in herbarium specimens of *A. artemisiifolia* in Europe as a function of collection day (Day), latitude of origin (Latitude) and collection year (Year). The predicted probability of observing flowers is plotted as a function of latitude (A), or collection day (B) for different collection years (mean collection year +/- 1 SD). The predicted probability of observing fruit is plotted against collection day for different collection years (mean collection year +/- 1 SD; C) or latitudes (mean collection latitude +/- 1 SD; D). Confidence intervals for the predictions are shown as are the raw data.
Figure S6. Distributions of $F_{ST}$ and $\theta_{HH}/\theta_{MM}$ 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 $\theta_{HH}/\theta_{MM}$ (dashed lines).
**Figure S7.** Distributions of $F_{ST}$ and $\theta_{nH}/\theta_{nM}$ 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 $\theta_{nH}/\theta_{nM}$ (dashed lines).
**Figure S8.** Distributions of $y_t$ values (calculated according to equation 1) for allele frequency shifts in 10,000 putatively neutral SNPs (histogram) between historic and modern samples from within 200km of Quebec City. Red line indicates the $y_t$ value for scaffold 2 position 56,025,662, a nonsynonymous SNP in *ELF3*. 
Figure S9. Haploblocks display extreme, divergent, local population structure (pale blue regions; first column). Haploblock regions (indicated by pale blue lines; second column) correspond to blocks of linkage disequilibrium (second highest $r^2$ in 0.5Mb windows) apparent using all modern samples (top triangle) but often reduced or absent using only samples homozygous for the more common haploblock genotype (bottom triangle). Haploblock genotypes were assigned by kmeans clustering (colours; third column) using the first two principal components of genetic variation across haploblock regions. Heterozygous haploblock genotypes show elevated mean per-site heterozygosity (fourth column; boxes denote mean and SEM).
Figure S10. Evidence of genotype-specific reductions in recombination in haploblock regions. Genetic distance (cM) against physical distance (bp) along a portion of each scaffold is shown. Haploblock regions are shown in pale blue. Example maps are displayed showing both low recombination rates (left) and high recombination rates (right) for each haploblock (scaffold 31 had few markers within the haploblock for most maps, and there was no clear evidence of substantial recombination within the haploblock for any of the maps). For scaffold 27, QTL-mapping data from Prapas et al.\textsuperscript{42} is shown: The marker closest to the large-effect flowering time QTL is represented by the red line. The 95% confidence intervals for male flowering time (dark purple), and height (light and dark purple) are indicated below the plots.
Figure S11. Genome-wide XtX scans between sampling locations within each range separately. Solid lines indicate Bonferroni-corrected significance derived from XtX p-values; dashed lines indicate the top 1% of genome-wide XtX values. Dark green highlights represent 10kb WZA windows with a top 5% WZA score in both ranges for XtX and at least one environmental variable. Pale blue highlights indicate haploblocks that overlap parallel WZA windows.
Figure S12. The distribution of slope estimates from generalized linear models of population allele counts against latitude for 10,000 randomly selected SNPs in each range. The vertical lines show the slope estimates for haploblocks with statistical associations with latitude in one range (table S12). Solid lines represent estimates in the 5% tail of each distribution while dotted lines fall below that cut-off.
Figure S13. Logistic regression models with 95% CI ribbons (see table S12-S17 for model details) of haploblock frequency (allele 1) against latitude for each haploblock across five time bins ranging from most historic (purple) to most modern (green).
**Figure S14.** The change in the slope of the relationship between latitude and haplotype frequency (see table S15) between historic and modern samples compared to the estimate of selection along the latitudinal cline for the seven haploblocks (estimated from modern data in each range). A strong relationship was detected in the invasive European, but not the native North American range.
**Figure S15.** HB27a region (ScBFxKa_27:18930440-23019032). XIX and XIX-EAA outlier windows zoomed from fig. 1C, and linkage disequilibrium ($r^2$) between SNPs in the region and HB27a haploblock genotype. A cluster of pectate lyase genes, consisting of the top BLAST hit for *Amba1* and closely-related paralogues, are indicated in red below.
Figure S16. The kmer coverage and model fit from GenomeScope 2.0 for 21mers from the PacBio HiFi and OmniC Illumina reads for *Ambrosia artemisiifolia*.
**Figure S17.** A smudgeplot for *Ambrosia artemisiifolia* used to estimate ploidy from heterzygous kmer pairs using 21mers from the HiFi reads.
Figure S18. A link density histogram, identified by proximity ligation sequencing. The x and y axes show mapping positions of the first and second read in read pairs.
Figure S19. Read depth and heterozygosity in 1Mb windows for Illumina reads used in the reference genome assembly mapped to the final version of the reference genome. Windows in the bottom 10% of read depth and heterozygosity values are indicated in red; windows overlapping haploblocks are indicated in pale blue.
Figure S20. A dot plot of an alignment of the reference *Ambrosia artemisiifolia* genome against itself.
**Figure S21.** Selection coefficients consistent with observed frequency changes of the HB448 haplotype, under models of positive selection (left) and balancing selection (right). The left-hand panel shows the distribution of $s$ values (gray) and the 95% CI for $s$ (in black: parallel to the x-axis), consistent with temporal change in the HB448 haploblock. The distribution of $s$ is based on simulations of 106 initial and final frequencies of the haploblock that are consistent with the estimated frequencies and error in the estimates. Eq. (S1) was used to calculate a value of $s$ for each set of initial and final frequencies during the time interval between 1883 (early historic) and 2014 (contemporary). The right-hand panel shows the selection coefficients under the balancing selection model that are consistent with observed changes in HB448 inversion frequencies in Europe. Across the range of possible polymorphic equilibrium frequencies under balancing selection, equilibria near unity (e.g., $p_{eq} = 0.99$) require modestly strong selection (average $S = 0.032$; 95% CI = [0.022, 0.043]) to explain the observed frequency changes in HB448, and lower equilibrium states (e.g., $p_{eq} < 0.90$) require stronger selection.
Figure S22. Simulated distribution of the scaled metric of divergence, $\hat{y}_t$, given different estimated values of the initial (historic) allele frequency. Each distribution is based on $10^4$ independent and neutrally evolving SNPs. The simulations use $n_0 = 182$, $n_t = 312$, and $t = 131$, with a moderate effective population size ($N_e = 10^4$). Histograms show the distribution of $\hat{y}_t$ estimates, the red circles show the simulated mean and 95% confidence intervals for simulated data, and the vertical broken red lines show +/- 1.96 SD, where SD is the square root of our analytical expression for $\text{var} [\hat{y}_t]$.
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. 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. 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:

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

(1)

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

We used eq. (1) to infer strengths of selection ($s$) that would be consistent with the estimated
change inversion frequencies over time. We used the estimated inversion frequencies across
two time points: (1) 0.4495 in the early historical samples with median sampling date of 1883
($n = 109$ diploid individuals), and (2) 0.7724 in a contemporary sample from 2014 ($n = 156$).

From these estimates, we simulated $10^6$ inversion frequencies per time point, each based on a
random sample from the binomial distribution with parameters $p$ and $2n$ corresponding to the
estimated frequency and sample size for the relevant time point (the factor of 2 in $2n$ accounts
for diploidy).

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

**Balancing selection model**

Although temporal changes in the HB448 haplotype are consistent with changes predicted
under the positive selection model presented above, we wished to also evaluate an alternative
model in which balancing selection favours evolution of the inversion towards an equilibrium
polymorphic state. To explore the strength of selection towards a hypothetical equilibrium within
the European range, we considered a simple model of overdominant selection. Note that the
overdominant selection model is dynamically equivalent to many other balancing selection
models provided the differences in fitness among genotypes are small (consistent with our
analysis below). Our results based on the overdominance model should, therefore, apply more broadly to other scenarios of balancing selection, including scenarios involving negative frequency-dependence and antagonistic pleiotropy\textsuperscript{83,84}.

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

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

where $s_1$ and $s_2$ refer to the fitness costs of being homozygous for inversion and standard haplotypes, respectively, and $p_{eq} = s_2(s_1 + s_2)^{-1}$ is the equilibrium frequency of the inversion.

Using a continuous-time approximation, we can solve for the overall selection coefficient, $S = s_1 + s_2$, that is consistent with a frequency shift from $p_0$ to $p_t < p_{eq}$ across $t$ generations:

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

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

Solving for $S$, gives us:

\[
S = \frac{p_{eq} log \left( \frac{1 - p_t}{1 - p_0} \right) + (1 - p_{eq}) log \left( \frac{p_t}{p_0} \right) - log \left( \frac{p_{eq} - p_t}{p_{eq} - p_0} \right)}{p_{eq}(1 - p_{eq})(t + log \left( \frac{1 - p_t}{1 - p_0} \right) + log \left( \frac{p_t}{p_0} \right) - log \left( \frac{p_{eq} - p_t}{p_{eq} - p_0} \right))}
\]

To infer the strength of selection ($S$) that would be consistent with observed inversion frequencies and a given equilibrium value ($p_{eq}$), we simulated $10^6$ inversion frequencies consistent with the estimated frequency and its sample size at historical time point (~1883) and $10^6$ frequencies consistent with the estimate for the contemporary sample. (Frequencies were simulated as described in the positive selection model section, above). We used each pair of simulated inversion frequencies and eq. (2) to infer the value of $S$ consistent with the frequency...
values. The resulting distribution of $10^6$ simulated $S$ values was used to calculate 95% CI for $S$ consistent with the data. We focused on equilibrium values outside of the 95% CI for contemporary inversion frequencies (i.e., values of $p_{eq}$ between 0.85 and 1). The results show that plausible selection coefficients under scenarios of balancing selection are consistently greater than those of the positive selection model (fig. S21). Selection under the positive selection model can, therefore, be regarded as a lower bound for the strength of selection consistent with the observed temporal changes in European HB448 inversion frequencies.
**S2: Selection estimated from spatial changes in haploblock frequency**

*Cline theory*

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

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

\[
\frac{dp(x)}{dt} = \frac{\sigma^2}{2} \frac{d^2 p(x)}{dx^2} + \Delta p_{set}(x)
\]

where \( \Delta p_{set}(x) \) is the local response to selection. With symmetrical strengths of selection at each side of the environmental transition, and no dominance, then \( \Delta p_{set}(x) \approx sp(x)(1 - p(x)) \) within the northern region of the range where the inversion is favoured, and \( \Delta p_{set}(x) \approx -sp(x)(1 - p(x)) \) 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 \( \sim 2s \) between inversion and standard haplotypes. Incorporating dominance does not change our results provided the dominance relations between the alleles are consistent across the range (i.e., under "parallel dominance"^88). At equilibrium between selection and migration, the maximum cline slope will be:
Following Roughgarden \(^{87}\), the equilibrium general solution for the cline is:

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

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

**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:

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

For steep clines—those with a narrow clinal region, where the maximum slope will be underestimated using the logit function—we have:
\[ d \log \left( \frac{p(x)}{1 - p(x)} \right) = \frac{2s}{\sqrt{\sigma^2}} \]

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 \( \beta_1 \) in table S12, we can calculate plausible ranges for the lower bound of \( \frac{\sqrt{s}}{\sigma} \) by multiplying the values for \( \beta_1 \) by \( \frac{\sqrt{3}}{4} \).

Comparisons of spatially varying selection among haploblocks

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

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

where \( A \) and \( B \) refer to different ranges or time points.
S3: A simple null model of temporal allele frequency changes under drift

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

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

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

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

\[
p_1 = \frac{x}{2N_e}
\]

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

\[
var(y_t) = \frac{var(p_t)}{tp_0(1 - p_0)} = \frac{1}{2N_e}
\]
In reality, error in the estimates of \( p_0 \) and \( p_t \) will also affect the test statistic, and this will tend to inflate the variance, but doesn’t alter the conclusion that (under a null model of drift with no selection) the distribution of the estimates of \( y_t \) (which we denote as \( \hat{y}_t \)) will be roughly independent of the initial allele frequencies in the historic sample. If we define \( \hat{p}_t \) and \( \hat{p}_0 \) is the estimates of the allele frequencies, then our test statistic is:

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

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

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

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

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

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

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

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

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

\[
\text{var}[\hat{p}_t|p_0] = \frac{E(p_t|p_0)(1 - E(p_t|p_0)) - \text{var}(p_t|p_0)}{n_t} + \text{var}(p_t|p_0) = \frac{p_0(1 - p_0)}{n_t} + \frac{tp_0(1 - p_0)(1 - \frac{1}{n_t})}{2N_e} \left(1 - \frac{1}{n_t}\right)
\]
where $N_e$ is the effective population size, and $n_0$ is the number of genes sampled in the historic population (e.g., for HB448, 91 individuals were sampled for the contemporary estimate in Europe; given diploidy, we have $n_0 = 182$).

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

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

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

$$\text{var} [\hat{p}_t|\hat{p}_0] = p_0(1 - p_0)\left[\frac{1}{n_t} + \frac{1}{2N_e} t \left(1 - \frac{1}{n_0}\right) \left(1 - \frac{1}{n_0}\right) + \frac{\hat{p}_0(1 - \hat{p}_0)}{n_0}\right]$$

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

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

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

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

Incidentally, the expression for $\text{var}[\hat{y}_t|\hat{p}_0]$ can be rearranged by solving for the effective population size across the $t$ generations, i.e.:
\[ N_e = \frac{1}{2} t \left( \frac{1}{n_t} \right) \left( 1 - \frac{1}{n_0} \right) \left( t \text{var}[\hat{y}_t] \right) - \frac{1}{n_0} - \frac{1}{n_t} \left( 1 - \frac{1}{n_0} \right) \]

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 \( \text{var}[\hat{y}_t] \).

**Simulations**

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


Ambrosia artemisiifolia L. in France from herbarium records. *J. Biogeogr.* 33, 665–673

13. van Boheemen, L. A. *et al.* Multiple introductions, admixture and bridgehead invasion
characterize the introduction history of Ambrosia artemisiifolia in Europe and Australia. *Mol.


Commun.* 8, 14435 (2017).


18. Orr, H. A. The population genetics of adaptation: The distribution of factors fixed during

19. Yeaman, S. & Whitlock, M. C. The genetic architecture of adaptation under migration-

20. Yeaman, S., Gerstein, A. C., Hodgins, K. A. & Whitlock, M. C. Quantifying how constraints


23. Wellenreuther, M. & Bernatchez, L. Eco-Evolutionary Genomics of Chromosomal


71. Cingolani, P. *et al.* 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. 6, 80–92 (2012).

72. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer