Massive haplotypes underlie ecotypic differentiation in sunflowers

Marco Todescov, Gregory L. Owens, Natalia Bercovich, Jean-Sébastien Légaré, Shaghayegh Soudi, Dylan O., Kaichi Huang, Katherine L. Ostevik, Emily B. M. Drummond, Ivana Imerovski, et al.

▶ **To cite this version:**

Marco Todescov, Gregory L. Owens, Natalia Bercovich, Jean-Sébastien Légaré, Shaghayegh Soudi, et al.. Massive haplotypes underlie ecotypic differentiation in sunflowers. Nature, In press, 10.1038/s41586-020-2467-6 . hal-02908080

HAL Id: hal-02908080
https://hal.inrae.fr/hal-02908080

Submitted on 28 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Article

Massive haplotypes underlie ecotypic differentiation in sunflowers

https://doi.org/10.1038/s41586-020-2467-6

Received: 28 September 2019

Accepted: 16 April 2020

Published online: 08 July 2020

Check for updates

Marco Todesco[1,2,11], Gregory L. Owens[1,2,3,11], Natalia Bercovich[1,2,11]✉, Jean-Sébastien Légaré[1,2,4,5], Shaghayegh Soudi[6], Dylan O. Burge[1,2], Kaichi Huang[1,2], Katherine L. Ostevik[7], Emily B. M. Drummond[1,2], Ivana Imerovski[1,2], Kathryn Lande[1,2], Mariana A. Pascual-Robles[1,2], Mihir Nanavati[4,10], Mojtaba Jahani[1,2], Winnie Cheung[1,2], S. Evan Staton[1,2], Stéphane Muños[8], Rasmus Nielsen[3], Lisa A. Donovan[9], John M. Burke[9], Sam Yeaman[6] & Loren H. Rieseberg[1,2]✉

Species often include multiple ecotypes that are adapted to different environments[1]. However, it is unclear how ecotypes arise and how their distinctive combinations of adaptive alleles are maintained despite hybridization with non-adapted populations[2–4]. Here, by resequencing 1,506 wild sunflowers from 3 species (*Helianthus annuus*, *Helianthus petiolaris* and *Helianthus argophyllus*), we identify 37 large (1–100 Mbp in size), non-recombining haplotype blocks that are associated with numerous ecologically relevant traits, as well as soil and climate characteristics. Limited recombination in these haplotype blocks keeps adaptive alleles together, and these regions differentiate sunflower ecotypes. For example, haplotype blocks control a 77-day difference in flowering between ecotypes of the silverleaf sunflower *H. argophyllus* (probably through deletion of a homologue of *FLOWERING LOCUS T* (*FT*)), and are associated with seed size, flowering time and soil fertility in dune-adapted sunflowers. These haplotypes are highly divergent, frequently associated with structural variants and often appear to represent introgressions from other—possibly now-extinct—congeners. These results highlight a pervasive role of structural variation in ecotypic adaptation.

Local adaptation is common in species that experience different environments across their range, often resulting in the formation of ecotypes—ecological races with distinct morphological and/or physiological characteristics that provide an environment-specific fitness advantage. Despite the prevalence of ecotypic differentiation, much remains to be understood about the genetic basis and evolutionary mechanisms that underlie its establishment and maintenance. In particular, a longstanding evolutionary question—dating to criticisms of Darwin's theories by his contemporaries[4]—concerns how such ecological divergence can occur when challenged by hybridization with non-adapted populations[2]. Local adaptation typically requires alleles at multiple loci that contribute to increased fitness in the same environment; however, different ecotypes are often geographically close and interfertile, and hybridization between them should break up adaptive allelic combinations[3].

To better understand the genetic basis of local adaptation and ecotypic differentiation, we conducted an in-depth study of genetic, phenotypic and environmental variation in three annual sunflower species, each of which includes multiple reproductively compatible ecotypes. Two species (*H. annuus* and *H. petiolaris*) have broad, overlapping distributions across North America. *Helianthus annuus*, the common sunflower, is generally found on mesic soils, but can grow in a variety of disturbed or extreme habitats, including semi-desert or frequently flooded areas. An especially well-characterized ecotype (formally known as *H. annuus* subsp. *texanus*) is adapted to the higher temperatures and herbivore pressures in Texas (USA)[5]. *Helianthus petiolaris*, the prairie sunflower, prefers sandier soils; ecotypes of this species are adapted to sand sheets and dunes[6]. The third species—*H. argophyllus*, the silverleaf sunflower—is endemic to southern Texas and includes both an early-flowering, coastal-island ecotype and a late-flowering inland ecotype[7].

Population structure of wild sunflowers

In a common garden experiment, we grew 10 plants from each of 151 populations of the 3 species, selected from across their native range (Fig. 1a); for each of these populations, we collected corresponding soil samples. We generated extensive records of developmental and morphological traits, and resequenced the genomes of 1,401 individual plants. We resequenced an additional 105 *H. annuus* plants to fill gaps in geographical coverage, as well as 12 outgroup taxa (Supplementary Table 1). Sunflower genomes are relatively large (*H. annuus*, 3.5 Gbp;

[1]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. [2]Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada. [3]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA. [4]Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada. [5]Data Science Institute, University of British Columbia, Vancouver, British Columbia, Canada. [6]Department of Biological Sciences, University of Calgary, Calgary, Alberta, Canada. [7]Department of Biology, Duke University, Durham, NC, USA. [8]LIPM, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France. [9]Department of Plant Biology, University of Georgia, Athens, GA, USA. [10]Present address: Microsoft Research, New York, NY, USA. [11]These authors contributed equally: Marco Todesco, Gregory L. Owens, Natalia Bercovich. ✉e-mail: nataliab@biodiversity.ubc.ca; lriesebe@mail.ubc.ca

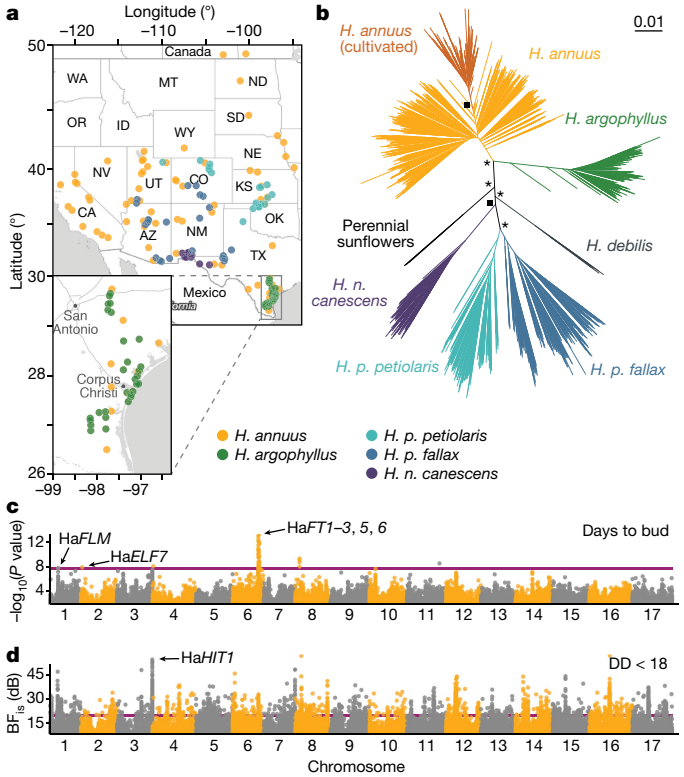


Fig. 1 | Population structure and association analyses of wild sunflowers. a, Map of wild sunflower populations surveyed in this study. **b**, Maximum-likelihood tree for the samples included in this study and previously described samples from cultivated sunflowers[16]. Bootstrap values for major nodes are reported (asterisks = 100; squares >89). **c**, Flowering-time GWA study for *H. annuus*. The purple line represents 5% Bonferroni-corrected significance. **d**, GEA analysis of degree-days below 18 °C (DD < 18) for *H. annuus*. The purple line represents a Bayes factor (BF_{is}) of 20 deciban (dB). Additional statistical information is provided in Methods.

H. petiolaris, 3.3 Gbp; and *H. argophyllus*, 4.3 Gbp[8]) and comprise >75% retrotransposon sequences[9]. We used enzymatic depletion[10] to reduce the proportion of repetitive sequences, which resulted in an average 6.34-fold coverage of gene space (median = 6.03) (Supplementary Table 1). We aligned sequencing reads to the reference genome of cultivated sunflower (a variety of *H. annuus*[9,11,12]), which resulted in sets of over four million high-quality single-nucleotide polymorphisms (SNPs) for each species (Extended Data Fig. 1).

A phylogeny based on these, and previously resequenced, sunflower samples is consistent with those of earlier studies[13,14]: *Helianthus annuus* and *H. argophyllus* are sister species, whereas *H. petiolaris* is placed in a separate clade. We found three separate lineages within our *H. petiolaris* collection, which correspond to the subspecies *H. p. fallax*, *H. p. petiolaris* and *H. p. canescens*. However, *H. p.* subsp. *canescens* falls within the *Helianthus niveus* clade, supporting an earlier classification[15]; owing to the smaller sample size (86 individuals), we omitted the *H. niveus canescens* clade from further analyses. Finally, dune-adapted ecotypes of *H. petiolaris* from Texas and Colorado (USA)[16] fall within *H. p. fallax* —despite the Texas populations being formally designated as the species *Helianthus neglectus*[17]—and we therefore analysed them as part of that clade (Fig. 1b).

Large haplotypes linked to adaptive traits

The large effective population size and outcrossing mating system of wild sunflowers[18] represent a major advantage for genome-wide association (GWA) studies, because the rapid decay of linkage disequilibrium permits mapping of phenotype–genotype associations to narrow genomic regions. GWA analyses of 87 traits identified numerous, strong links between phenotypic variation and regions of the sunflower genome (Supplementary Table 2). For example, we observed extensive variation in flowering time for all three species (Extended Data Fig. 2a), consistent with its fundamental role in sunflower adaptation (and that of plants more generally)[19,20]. For *H. annuus*, significant associations were found with the sunflower homologues of known regulators of flowering time, including *FT*[21], *FLOWERING LOCUS M* (*FLM*)[22] and *EARLY FLOWERING 7* (*ELF7*)[23] (Fig. 1c). We also identified genomic regions that are strongly associated with environmental and soil variables in genotype–environment association (GEA) analyses, which suggests a role in adaptation to particular habitats (Supplementary Table 2). For example, several temperature-related variables showed strong associations with the sunflower homologue of *HEAT-INTOLERANT 1* (*HIT1*), which mediates resistance to heat stress by regulating plasma membrane thermotolerance in *Arabidopsis thaliana*[24] (Fig. 1d).

In several cases, GWA and GEA signals spanned very large regions of the genome for traits that are known to be important for local adaptation, and to differentiate ecotypes in sunflower. A particularly notable GWA plateau occurred between coastal-island and inland populations of *H. argophyllus*. Inland populations flower late in summer and can grow extremely tall (>4 m), whereas shorter, early-flowering individuals occur at high frequency on the barrier islands of the Gulf of Mexico (Fig. 2a, b). Selection experiments indicate that late flowering in the interior is favoured[7], presumably to avoid flowering during the extremely hot and dry summer, whereas early flowering appears to be advantageous under less-harsh conditions on the barrier islands. Our flowering-time GWA analyses in *H. argophyllus* identified a single, highly significant association that spans about 30 Mbp on chromosome 6 (Fig. 2c, d), and which is also associated with leaf nitrogen and carbon content (Extended Data Fig. 2b). Principal component analysis (PCA) of this region suggested the presence of two main haplotypes, with intermediate individuals being heterozygotes (Fig. 2e). We extracted haplotype-informative sites and visualized ancestry across the region, which revealed that recombination was very limited. A 10-Mbp region (130–140 Mbp) is perfectly correlated with flowering-time phenotypes and explains 88.2% of variance in days to bud (Fig. 2f). The early-flowering haplotype acts dominantly; plants that carry at least one copy of it flower, on average, 77 days earlier than late-flowering plants (Fig. 2g). This region contains five of the six sunflower homologues of the flowering-time regulator *FT* (Fig. 2f). The GWA signal drops sharply around the *H. argophyllus* (Ha)*FT1* locus (Fig. 2d), which underlies differences in photoperiodic responses between wild and cultivated sunflower[25]. Analysis of an unfiltered SNP dataset revealed that this pattern is due to the absence of reads that map to the region in plants carrying the late-flowering haplotype (only SNPs with data for >90% of individuals were used for GWA studies). This is consistent with the presence of one or more deletions—including the Ha*FT1* locus—in late-flowering *H. argophyllus* (Fig. 2h, Extended Data Fig. 2c). Accordingly, the Ha*FT1* sequence cannot be amplified from genomic DNA from late-flowering plants, and no Ha*FT1* expression is detected in these plants (Fig. 2i, j, Extended Data Fig. 2d). Expression of Ha*FT1* could instead be detected in early-flowering plants (Fig. 2j, Extended Data Fig. 2d), and transgenic introduction of this Ha*FT1* allele restored early flowering in the otherwise late-flowering *A. thaliana ft-10* mutant (Fig. 2k, l). To explore the origins of these haplotypes, we constructed a phylogeny of the non-recombined 10-Mbp region in chromosome 6 (Fig. 2m). We found that the two haplotypes are highly divergent, and that the early-flowering haplotype was introgressed from *H. annuus* (D statistic = 0.844 ± 0.006, $P < 10^{-20}$, two-sided) (Fig. 2g). Although a role of the other homologues of *FT* (Extended Data Fig. 2e–g) or other genes in the region cannot be excluded, these results strongly suggest that introgression of a functional Ha*FT1* copy from *H. annuus* was key in the establishment of early-flowering *H. argophyllus*.

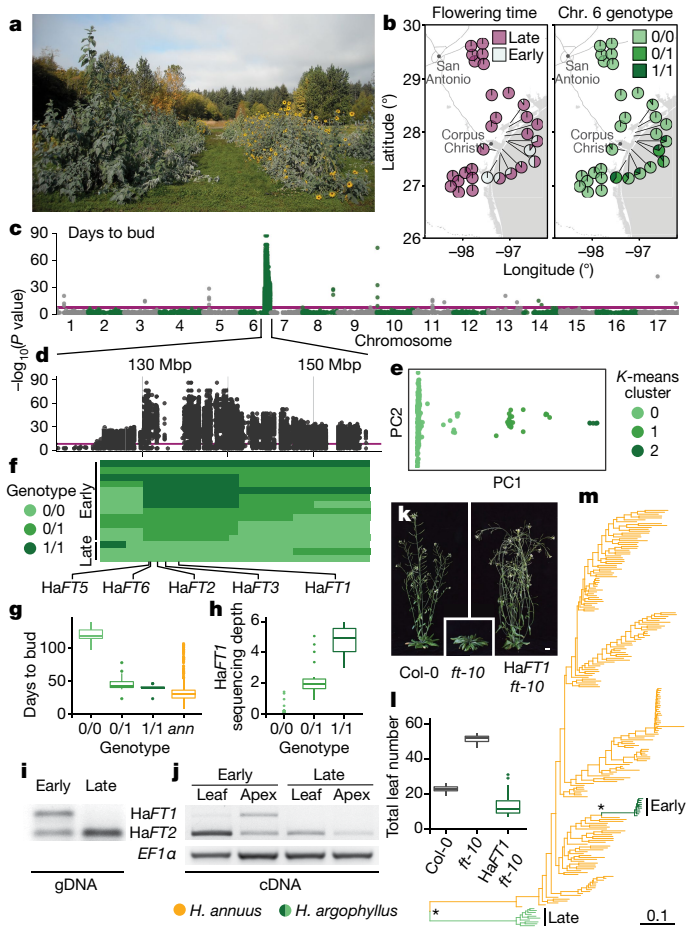


Fig. 2 | A large introgression from *H. annuus* containing a functional Ha*FT1* gene causes early flowering in coastal *H. argophyllus*. **a**, Inland (left) and coastal-island (right) *H. argophyllus* plants. Image provided by B. T. Moyers. **b**, Distribution of late- and early-flowering ecotypes (left), and of haplotypes of the causal 10-Mbp region on chromosome 6 (right). **c**, **d**, Flowering-time GWA analysis in *H. argophyllus* (**c**) and enlarged view of the bottom of chromosome 6 (**d**). The purple line represents 5% Bonferroni-corrected significance. **e**, PCA of the last 30 Mbp of chromosome 6. Three clusters are defined by principal component 1 (PC1). **f**, Schematic of all unique haplotypes found at the bottom of chromosome 6, and corresponding flowering time. Chromosome positions match **d**. **g**, Flowering times associated with different genotypes at the approximately 130–140-Mbp region of chromosome 6 (*ann*, *H. annuus*). **h**, Sequencing depth of SNPs in the Ha*FT1* gene. **i**, PCR on genomic DNA from early- and late-flowering *H. argophyllus* plants. **j**, Expression analysis in mature leaves or shoot apices of the plants examined in **i**, grown for six weeks in long-day conditions (14 h light; 10 h dark). Cleaved-amplified polymorphic sequence (CAPS) markers were used to distinguish Ha*FT1* from Ha*FT2*. For gel source data, see Supplementary Fig. 1. **k**, Six-week-old *A. thaliana* plants grown in long-day conditions. *ft-10* is a late-flowering *A. thaliana FT* mutant in Col-0 background. Scale bar, 1 cm. **l**, Flowering time (as total leaf number) for primary transformants that express the *FT1* gene from *H. argophyllus* in the *ft-10* background. Difference between *ft-10* and Ha*FT1 ft-10* is significant ($P < 10^{-8}$). **m**, Maximum likelihood phylogeny of the 130–140-Mbp region on chromosome 6 in *H. argophyllus* and *H. annuus*. Bootstrap values for major nodes are reported (asterisks = 100). Additional statistical information is provided in Methods.

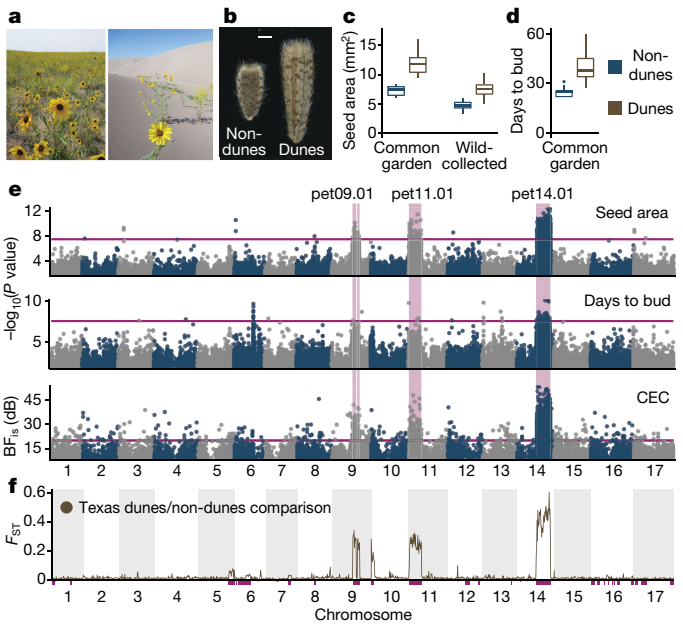


Fig. 3 | Large non-recombining haplotypes control dune adaptation in *H. p. fallax*. **a**, Sand sheet (left) or sand dune (right) populations of *H. petiolaris*. The dune sunflowers image was provided by J. D. Herndon. **b**, Representative seeds from dune-adapted and non-dune-adapted plants grown in a common garden. Scale bar, 1 mm. **c**, Seed size (area), averaged for eight seeds per plant. Seeds were collected in the wild or from a common garden. In both conditions, seeds from dune-adapted plants are about 60% larger than those of non-dune-adapted plants. $P = 2.2 \times 10^{-16}$ for wild plants, $P = 1.2 \times 10^{-6}$ for common garden plants. **d**, Flowering time in a common garden. $P = 8.1 \times 10^{-8}$. **e**, Seed-size and flowering-time GWA analyses, and CEC (soil fertility) GEA analysis for *H. p. fallax*. Purple lines represent 5% Bonferroni-corrected significance (in the GWA analyses), and $BF_{is} = 20$ dB (in the GEA analysis). Haploblock predictions corresponding to three significant plateaus are highlighted in purple. **f**, F_{ST} values in 2-Mbp non-overlapping sliding windows for comparisons between dune- and non-dune-adapted Texas populations of *H. p. fallax*. Purple bars represent predicted haploblocks. Some predicted haploblocks are fragmented owing to rearrangements in *H. petiolaris* relative to the *H. annuus* reference genome. Additional statistical information is provided in Methods.

We found another example of GWA and GEA plateaus that underlie ecotypic differentiation in *H. petiolaris*, which has repeatedly adapted to sand dunes in Texas and Colorado[6]. Dune populations exhibit distinctive phenotypes compared to populations that grow close to the same dunes (Fig. 3a–d), the most notable of which are seed size and length (Fig. 3b, c); large seeds confer a strong fitness advantage on sand dunes[6], possibly by providing seedlings with enough resources to emerge after burial by sand. Dunes also are low in nutrients, and dune sunflowers use soil nutrients more efficiently than their non-dune counterparts[26]. Our GWA analyses for seed size and flowering time, and our GEA analyses of soil characteristics (including cation exchange capacity (CEC), a measure of soil fertility), in *H. p. fallax* identified three multi-Mbp regions on chromosomes 9, 11 and 14 (Fig. 3d, e, Extended Data Fig. 3a, b). All three regions are highly differentiated between dune and non-dune populations from Texas (Fig. 3f), and two of the three regions differentiate dune and non-dune populations in Colorado[27] (Extended Data Fig. 3c), which suggests a fundamental role in maintaining the dune ecotype. Although strong differentiation in dune populations could confound these associations, colocalization of the plateaus on chromosomes 11 and 14 with known quantitative trait loci for seed size that differentiated dune and non-dune populations[28], coupled with the observation of weaker associations with flowering time in *H. p. petiolaris* for the chromosome-9 and -11 regions (Extended Data Fig. 3d), further confirm a direct role of these regions in controlling dune-specific traits.

Highly divergent haploblocks are common

The identification of these GWA and GEA plateaus suggests a broader role of large, non-recombining haplotype blocks (hereafter 'haploblocks') in adaptation. Therefore, we used a local PCA approach to

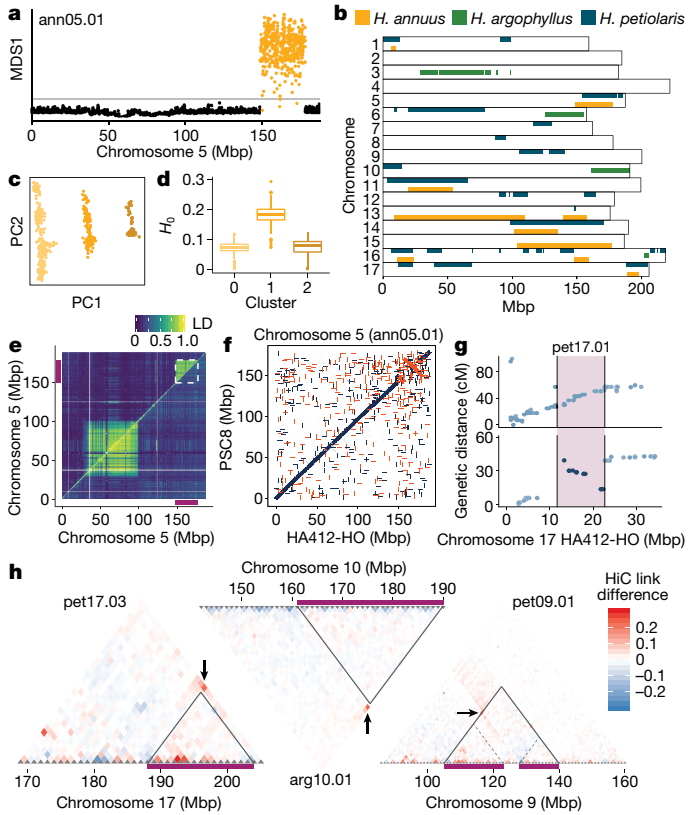


Fig. 4 | Large haploblocks are pervasive in wild sunflowers and are associated with structural variants. a, Local PCA output for putative segregating haploblock (ann05.01). Dots represent 100 SNP windows; outlier windows are highlighted. MDS, multidimensional scaling. **b**, Genomic positions of identified haploblocks. Some predicted haploblocks are fragmented owing to rearrangements relative to the *H. annuus* reference genome. **c**, PCA for the ann05.01 region in *H. annuus*. **d**, Heterozygosity at ann05.01 for the three clusters of plants identified in **c**. **e**, Linkage disequilibrium (LD) plot for chromosome 5. Top triangle, all *H. annuus* plants; bottom triangle, plants homozygous for the more-common haploblock allele. Colours represent the second highest R^2 value in 0.5-Mbp windows. Purple bars and the white box represent ann05.01. **f**, Comparison of chromosome-5 organization between two sunflower reference assemblies. Red, reverse; blue, forward. **g**, Comparison between genetic maps for *H. p. fallax* and the chromosome-17 HA412-HOv2 reference sequence. A predicted haploblock (pet17.01) is highlighted in purple. **h**, Comparison of HiC link patterns between pairs of plants with different haplotypes at three haploblocks. Red and blue dots show increased and decreased long-distance interactions, respectively, in one sample, consistent with structural variants. Differences highlighted by black arrows are in the >99.9th percentile compared to all other possible interactions at the same distance across the genome. Purple bars and solid black lines represent predicted haploblocks. The dune haplotype for pet09.01 is split in two fragments (dotted lines) when aligned to the *H. annuus* reference genome. Additional statistical information is provided in Methods.

identify other large genomic regions with distinct population structure[29] (Fig. 4a). Across the 3 species, we found 37 such regions, which range from 1 to 100 Mbp in size and represent 4–16% of the genome (Fig. 4b, Extended Data Table 1). These haploblocks are characterized by high linkage disequilibrium; PCAs in the haploblock regions separated individual genotypes into three clusters, with the middle cluster having higher heterozygosity (Fig. 4c–e, Extended Data Figs. 4, 5). This is consistent with the two extreme clusters representing plants homozygous for two distinct haplotypes, and the middle cluster representing heterozygotes. No, or very little, recombination is observed between haplotypes, but generally no reduction in recombination is found within haplotypes (Fig. 4e, Extended Data Figs. 4, 5).

These patterns match expectations for large, segregating structural variants. Theory indicates that structural variants can facilitate adaptive divergence in the face of gene flow by reducing recombination between locally adaptive alleles[30–32]. In particular, inversions have previously been shown to control adaptive phenotypic variation (for example, migration[33], colour[34], flowering time[35] or adaptation to altitude[36]), and to be associated with environmental clines[37]. We used three approaches to determine whether these haploblocks are associated with structural variants (Extended Data Table 1). First, we compared the genome assemblies of two cultivars of *H. annuus* that have opposite genotypes at haploblock regions on chromosomes 1 and 5 (designated ann01.01 and ann05.01, respectively). We found one and two large inversions, respectively, at these regions (Fig. 4f, Extended Data Fig. 6a). We also aligned ten *H. annuus* and four *H. petiolaris* genetic maps to the sunflower reference genome; we observed suppressed recombination at ten haploblocks, and evidence for three haploblocks being caused by large inversions (Fig. 4g, Extended Data Fig. 6b, c). Finally, we used chromosome conformation capture sequencing (HiC)[38] to compare pairs of early- and late-flowering *H. argophyllus* and dune- and non-dune-adapted *H. petiolaris*, and looked for differences in physical linkage at haploblock regions. We found support for structural variants—ranging from likely full-length inversions to more-complex rearrangements—at 11 regions in *H. petiolaris* and one in *H. argophyllus* (Fig. 4h, Extended Data Fig. 7). For one haploblock for each species, we could find no evidence of structural variants in our HiC data, which suggests that recombination might be suppressed by other mechanisms in these regions. We also confirmed that large structural variants underlie four of the haploblocks detected in wild *H. annuus* by comparing our HiC data to those for the HA412-HO reference cultivar (a version of *H. annuus*) (Extended Data Fig. 7). These results point to structural variants being associated with most of the haploblocks that we detected.

Of the 37 haploblocks we identified, two (arg06.01 and arg06.02) correspond to the chromosome-6 region that is associated with flowering time in *H. argophyllus*, and three (pet09.01, pet11.01 and pet14.01) correspond with seed size, flowering time and CEC plateaus in *H. petiolaris* (Fig. 4b, Extended Data Table 1). We also identified four additional haploblocks that colocalize with regions of high genetic differentiation between dune-adapted and non-dune-adapted ecotypes of *H. petiolaris* (Fig. 3f, Extended Data Fig. 3c, e), which bring the total number of haploblocks associated with dune adaptation to seven—four of which are shared between both independent dune ecotypes (that is, Texas and Colorado) (Extended Data Fig. 5).

Our phylogenetic analysis found that these dune-adapation-associated haploblocks predate the split between *H. p. fallax* and *H. p. petiolaris*, and that five of the haploblocks are polymorphic in both subspecies (Fig. 5b, Extended Data Fig. 3f). Such high levels of divergence are common to most haploblock regions (Fig. 5a). For the two haploblocks that are polymorphic between the *H. annuus* reference genomes (that is, ann01.01 and ann05.01) (Fig. 4e, f, Extended Data Fig. 6a), sequence identity between haplotypes is 94–95%—much lower than the 99.4% for the rest of the genome. Divergence times between all but one of the haploblocks exceed 1 million years, and in most cases (32 out of 37) predate the *H. annuus*–*H. argophyllus* speciation event[39] (Fig. 5b). This seems at odds with the observation that haploblock polymorphisms are not shared between sunflower species. Ancient haploblocks could have been maintained in selected lineages (possibly by balancing selection[40]), but this should result in transpecific polymorphisms. Alternatively, the haploblocks could be more recently introgressed from divergent taxa[41]; this hypothesis is supported for four of the *H. argophyllus* haploblocks, in which one haplotype is phylogenetically closer to *H. annuus* than to *H. argophyllus* (Fig. 2m). However, a donor species could not be identified for more-divergent haploblocks, which raises the possibility that these haploblocks may be introgressed from one or more now-extinct taxa.

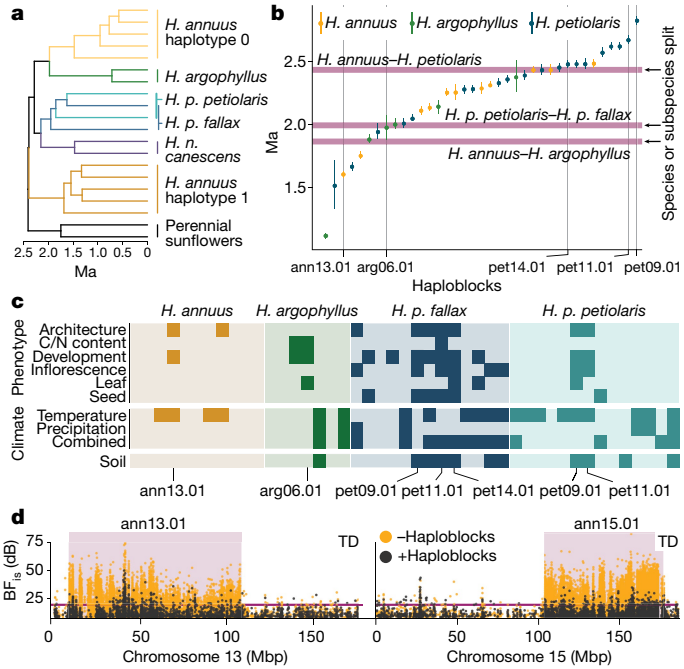


Fig. 5 | Haploblocks are highly divergent and are associated with multiple ecologically relevant traits and environmental variables. a, Bayesian phylogeny of ann05.01. Ma, million years ago. **b**, Divergence-time estimates for haploblocks, relative to those of different species or subspecies of wild sunflowers. **c**, Summary of GWA and GEA analyses for haploblocks treated as single loci. Darker squares represent categories that contain at least one trait or variable that is significantly associated with a haploblock ($P < 0.001$; $BF_{is} > 10$). **d**, GEA analyses for temperature difference (TD) (a measure of continentality) using a kinship matrix and PCA covariate including (black dots) or excluding (yellow dots) the haploblock regions. Haploblocks are highlighted in purple. Purple lines represent $BF_{is} = 20$ dB (for the GEAs). Additional statistical information is provided in Methods.

Haploblocks underlie ecotype divergence

As we have shown, haploblocks can have strong associations with phenotypic traits and environmental variables (Figs. 2c, 3e, Extended Data Figs. 2b, 3b, d), but these examples represent only a small proportion of the total haploblock regions that we identified. We therefore considered whether the other haploblocks are also involved in local adaptation. Theory suggests that structural variants are likely to establish by capturing multiple adaptive alleles[30]; consistent with this, when we treated haploblocks as individual loci, we found that haploblocks are often associated with multiple types of trait (Fig. 5c, Extended Data Figs. 8, 9).

Some of the strongest associations we identified with this approach did not appear in our initial GWA and GEA analyses. Haploblocks are large enough to affect the genome-wide estimates of relatedness between individuals (kinship and PCA) routinely used to compensate for population structure in GWA and GEA analyses, which can result in their association signal being masked[42]. This is particularly evident for ann13.01, which—at about 100 Mbp—is the largest of the haploblocks we identified; significant plateaus for temperature difference (a measure of climate continentality) and flowering time are revealed only once haploblock regions are removed from the kinship covariate (Fig. 5d, Extended Data Fig. 10a, b). This haploblock, and several others, appear to differentiate Texas populations of *H. annuus* from the rest of the range (Extended Data Figs. 5, 10c), consistent with the distribution of the *texanus* ecotype of *H. annuus*[43]. Similar to the comparisons for dune adaptation in *H. petiolaris* (Extended Data Fig. 3e), haploblocks are more differentiated than SNPs in comparisons between Texas and other populations ($t(10) = 4.01$, $P = 0.0024$, two-sided t-test) (Extended Data Fig. 10d), which supports a role for haploblocks in the local adaptation of this subspecies, or in increasing its reproductive isolation with a local congener[44].

Conclusions

We have identified numerous highly divergent, multi-Mbp-long haploblocks in wild sunflowers, many of which appear to underlie ecotype formation: four in the early-flowering ecotype of *H. argophyllus*; seven in the *texanus* ecotype of *H. annuus*; and seven in dune-adapted ecotypes of *H. petiolaris* (Extended Data Fig. 5). These haploblocks are often linked to large structural variants (especially inversions), which provide a straightforward mechanism for suppressing recombination between haplotypes and thereby maintaining adaptive allelic combinations. The total number and effects of such haplotypes are probably even larger, as our approach is biased towards the detection of divergent and large (>1 Mbp) haploblocks.

Ecotypic differentiation is often seen as a first step towards the generation of new species[1], and the ecotypes discussed here appear to represent different stages in the speciation continuum. The coastal-island ecotype of *H. argophyllus* is the least divergent and the only known reproductive barrier with the inland ecotype is flowering time[7], which provides only modest protection from gene flow. By contrast, multiple reproductive barriers differentiate the dune-adapted ecotypes of *H. p. fallax* from nearby non-dune-adapted populations[6,28,45], reducing—but not eliminating—gene flow[17,46]. Notably, several haploblocks are associated both with traits favouring local adaptation and with those contributing to reproductive isolation (for example, seed size and flowering time, respectively, in the dune ecotypes); this architecture facilitates speciation with gene flow[31,32]. More generally, flowering time mapped to one or more haploblocks in all ecotypes, which suggests that it has an especially important role in ecotype formation—perhaps owing to its dual role in local adaptation and assortative mating[47]. Because our common garden plants were grown from wild-collected seeds, trait variation might be affected by environmental maternal effects. However, the strong GWA and GEA signals observed indicate a sizable genetic component to this variation.

An unanswered question is how the linked combinations of locally favoured mutations found in haploblocks arose. It is possible that sets of locally adaptive alleles initially developed in geographically isolated populations[48]. Secondary contact and hybridization would favour the evolution of reduced recombination among such alleles through the establishment of structural variants[30] or other recombination modifiers[31]. An origin through introgression would also help to account for the high divergence and massive size of many haploblocks, as well as the lack of shared haploblock polymorphisms between species. After haploblock establishment, new locally adaptive mutations would be more likely to persist under migration–selection balance if linked to other adaptive alleles[49], potentially leading to the outsized effects reported here. Our work reveals a modular genetic architecture that underlies ecotype formation, an unforeseen origin of many locally adapted gene modules through introgression and a critical role of recombination modifiers—especially structural variants—in adaptive divergence with gene flow.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2467-6.

1. Clausen, J. *Stages in the Evolution of Plant Species* (Cornell Univ. Press, 1951).
2. Endler, J. A. Gene flow and population differentiation. *Science* **179**, 243–250 (1973).

3. Felsenstein, J. Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution* **35**, 124–138 (1981).
4. Romanes, G. J. Physiological selection; an additional suggestion on the origin of species. *Zool. J. Linn. Soc.* **19**, 337–411 (1886).
5. Whitney, K. D., Randell, R. A. & Rieseberg, L. H. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytol.* **187**, 230–239 (2010).
6. Ostevik, K. L., Andrew, R. L., Otto, S. P. & Rieseberg, L. H. Multiple reproductive barriers separate recently diverged sunflower ecotypes. *Evolution* **70**, 2322–2335 (2016).
7. Moyers, B. T. *The Landscape of Divergence in Silverleaf Sunflowers*. PhD thesis, Univ. of British Columbia (2015).
8. Qiu, F. et al. Phylogenetic trends and environmental correlates of nuclear genome size variation in *Helianthus* sunflowers. *New Phytol.* **221**, 1609–1618 (2019).
9. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
10. Shagina, I. et al. Normalization of genomic DNA using duplex-specific nuclease. *Biotechniques* **48**, 455–459 (2010).
11. Staton, S. E. & Rieseberg, L. H. Sunflower Genome Database, https://www.sunflowergenome.org/ (2019).
12. INRA. INRA Sunflower Bioinformatics Resources, https://www.heliagene.org/ (2019).
13. Baute, G. J., Owens, G. L., Bock, D. G. & Rieseberg, L. H. Genome-wide genotyping-by-sequencing data provide a high-resolution view of wild *Helianthus* diversity, genetic structure, and interspecies gene flow. *Am. J. Bot.* **103**, 2170–2177 (2016).
14. Stephens, J. D., Rogers, W. L., Mason, C. M., Donovan, L. A. & Malmberg, R. L. Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *Am. J. Bot.* **102**, 910–920 (2015).
15. Heiser, C. B. & Smith, D. M. *The North American Sunflowers (Helianthus)* (Seeman Printery, 1969).
16. Hübner, S. et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants* **5**, 54–62 (2019).
17. Raduski, A. R., Rieseberg, L. H. & Strasburg, J. L. Effective population size, gene flow, and species status in a narrow endemic sunflower, *Helianthus neglectus*, compared to its widespread sister species, *H. petiolaris*. *Int. J. Mol. Sci.* **11**, 492–506 (2010).
18. Strasburg, J. L. & Rieseberg, L. H. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*—large effective population sizes and rates of long-term gene flow. *Evolution* **62**, 1936–1950 (2008).
19. Blackman, B. K., Michaels, S. D. & Rieseberg, L. H. Connecting the sun to flowering in sunflower adaptation. *Mol. Ecol.* **20**, 3503–3512 (2011).
20. Zan, Y. & Carlborg, Ö. A polygenic genetic architecture of flowering time in the worldwide *Arabidopsis thaliana* population. *Mol. Biol. Evol.* **36**, 141–154 (2019).
21. Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M. & Araki, T. A pair of related genes with antagonistic roles in mediating flowering signals. *Science* **286**, 1960–1962 (1999).
22. Werner, J. D. et al. Quantitative trait locus mapping and DNA array hybridization identify an *FLM* deletion as a cause for natural flowering-time variation. *Proc. Natl Acad. Sci. USA* **102**, 2460–2465 (2005).
23. Cao, Y., Wen, L., Wang, Z. & Ma, L. SKIP interacts with the Paf1 complex to regulate flowering via the activation of *FLC* transcription in *Arabidopsis*. *Mol. Plant* **8**, 1816–1819 (2015).
24. Wang, L. C. et al. Involvement of the *Arabidopsis* HIT1/AtVPS53 tethering protein homologue in the acclimation of the plasma membrane to heat stress. *J. Exp. Bot.* **62**, 3609–3620 (2011).
25. Blackman, B. K. et al. Contributions of flowering time genes to sunflower domestication and improvement. *Genetics* **187**, 271–287 (2011).
26. Brouillette, L. C. & Donovan, L. A. Nitrogen stress response of a hybrid species: a gene expression study. *Ann. Bot.* **107**, 101–108 (2011).
27. Andrew, R. L. & Rieseberg, L. H. Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution* **67**, 2468–2482 (2013).
28. Ostevik, K. L. *The Ecology and Genetics of Adaptation and Speciation in Dune Sunflowers*. PhD thesis, Univ. of British Columbia (2016).
29. Li, H. & Ralph, P. Local PCA shows how the effect of population structure differs along the genome. *Genetics* **211**, 289–304 (2019).
30. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
31. Ortiz-Barrientos, D., Engelstädter, J. & Rieseberg, L. H. Recombination rate evolution and the origin of species. *Trends Ecol. Evol.* **31**, 226–236 (2016).
32. Trickett, A. J. & Butlin, R. K. Recombination suppressors and the evolution of new species. *Heredity* **73**, 339–345 (1994).
33. Arostegui, M. C., Quinn, T. P., Seeb, L. W., Seeb, J. E. & McKinney, G. J. Retention of a chromosomal inversion from an anadromous ancestor provides the genetic basis for alternative freshwater ecotypes in rainbow trout. *Mol. Ecol.* **28**, 1412–1427 (2019).
34. Joron, M. et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
35. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
36. Fustier, M. A. et al. Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude. *PLoS Genet.* **15**, e1008512 (2019).
37. Wellenreuther, M., Rosenquist, H., Jaksons, P. & Larson, K. W. Local adaptation along an environmental cline in a species with an inversion polymorphism. *J. Evol. Biol.* **30**, 1068–1077 (2017).
38. Belton, J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
39. Mason, C. M. How old are sunflowers? A molecular clock analysis of key divergences in the origin and diversification of *Helianthus* (Asteraceae). *Int. J. Plant Sci.* **179**, 182–191 (2018).
40. Wellenreuther, M. & Bernatchez, L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
41. Jay, P. et al. Supergene evolution triggered by the introgression of a chromosomal inversion. *Curr. Biol.* **28**, 1839–1845. (2018).
42. Lotterhos, K. E. The effect of neutral recombination variation on genome scans for selection. *G3* **9**, 1851–1867 (2019).
43. Heiser, C. B., Jr. Hybridization in the annual sunflowers: *Helianthus annuus* × *H. debilis* var. *cucumerifolius*. *Evolution* **5**, 42–51 (1951).
44. Hooper, D. M. & Price, T. D. Chromosomal inversion differences correlate with range overlap in passerine birds. *Nat. Ecol. Evol.* **1**, 1526–1534 (2017).
45. Heiser, C. B. Three new annual sunflowers (*Helianthus*) from the southwestern United States. *Rhodora* **60**, 272–283 (1958).
46. Andrew, R. L., Kane, N. C., Baute, G. J., Grassa, C. J. & Rieseberg, L. H. Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Mol. Ecol.* **22**, 799–813 (2013).
47. Kirkpatrick, M. Reinforcement and divergence under assortative mating. *Proc. R. Soc. Lond. B* **267**, 1649–1655 (2000).
48. Feder, J. L., Gejji, R., Powell, T. H. & Nosil, P. Adaptive chromosomal divergence driven by mixed geographic mode of evolution. *Evolution* **65**, 2157–2170 (2011).
49. Yeaman, S. & Whitlock, M. C. The genetic architecture of adaptation under migration–selection balance. *Evolution* **65**, 1897–1911 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Investigators were not blinded to sample identity during experiments.

Seed and soil collection

During the summer of 2015 we visited 192 wild populations spanning the native distributions of *H. annuus*, *H. petiolaris* and *H. argophyllus*, and collected seeds from 21–37 individuals from each population. Seeds from ten additional populations of *H. annuus* had been previously collected in the summer of 2011. Three to five soil samples (0–25-cm depth) were collected with a soil corer at each population, from across the area in which seeds were collected. Soils were air dried in the field, further dried at 60 °C in to the lab, and passed through a 2-mm sieve to remove roots and rocks. Soils were then submitted to Midwest Laboratories for analysis.

Common garden

Ten mother plants were randomly selected from each of 151 populations that were included in the common garden experiment. Ten seeds from each of these plants were surface-sterilized by immersing them for 10 min in a 1.5% sodium hypochlorite solution. Seeds were then rinsed twice in distilled water and treated for at least one hour in a solution of 1% PPM (Plant Cell Technologies), a broad-spectrum biocide–fungicide, to minimize contamination, and 0.05 mM gibberellic acid (Sigma-Aldrich). The seeds were then scarified, de-hulled, and kept for two weeks at 4 °C in the dark on filter paper imbibed with a 1% PPM solution. Following this, seeds were kept in the dark at room temperature until they germinated, and then transplanted in peat pots. Seedlings were grown in a greenhouse for two weeks and then moved to an open-sided greenhouse for a week for acclimatization. Plants were transplanted into three separate fields (one for each sunflower species) at the Totem Plant Science Field Station of the University of British Columbia on the 25 May 2016 (*H. argophyllus*), 2 June 2016 (*H. petiolaris*) and 7 June 2016 (*H. annuus*). Within each field, pairs of plants from the same population of origin were sown using a completely randomized design. At least three flowers from each plant were bagged before anthesis to prevent pollination, and manually crossed to an individual from the same population of origin. Phenotypic measurements were performed throughout plant growth, and leaves, stem, inflorescences and seeds were collected and digitally imaged to extract relevant morphometric data using Fiji[50,51] and Tomato Analyzer[52] (Supplementary Table 1). Plants were grown until the beginning of November, by which point almost all the plants had flowered.

DNA isolation, library preparation and sequencing

Tissue from young leaves was collected from all individual plants, and genomic DNA was extracted from leaf tissue using a CTAB protocol (modified from refs. [53,54]), the DNeasy Plant Mini Kit or a DNeasy 96 Plant Kit (Qiagen). DNA was sheared to an average fragment size of 400 bp using a Covaris M220 ultrasonicator (Covaris), following the manufacturer's recommendations. Seven hundred and fifty ng of sheared DNA were used as starting material to prepare paired-end whole-genome shotgun (WGS) Illumina libraries for 719 *H. annuus*, 488 *H. petiolaris* and 299 *H. argophyllus* individuals, and 12 additional samples from annual and perennial sunflowers (Supplementary Table 1), using a protocol largely based on ref. [55], the TruSeq DNA Sample Preparation Guide from Illumina (Illumina) and ref. [56]. End-repairing of the sheared DNA fragments was performed using the NEBNext End Repair Module (NEB). The fragments were then A-tailed using Klenow Fragment (3′→5′exo-; NEB) and ligated to 24-bp-long, non-barcoded adapters with a 3′ T-overhang using the Quick Ligation Kit (NEB). After each enzymatic step, the reactions were purified using 1.6 volumes of paramagnetic SPRI beads, prepared according to ref. [56]. An enrichment step was then performed using KAPA HiFi HotStart ReadyMix (Roche) and short, non-indexed primers that do not extend the adapters. The reactions were then purified using 1.6 volumes of SPRI beads.

To reduce the proportion of repetitive sequences, libraries were treated with a Duplex-Specific Nuclease (DSN; Evrogen), following the protocols reported in refs. [10,57], with modifications. Depletion conditions were optimized for the sunflower genome by quantitative PCR; relative abundance of chloroplast DNA and transposable elements before and after depletion was estimated using a primer pair recognizing a chloroplast gene, and degenerate primers recognizing one of the most abundant transposon families in the sunflower genome, and comparing them to the abundance of the single copy Ha*LFY* gene. Libraries were concentrated to 160 ng/µl using SPRI beads. Three µl of libraries were mixed to 1 µl of hybridization buffer (200 mM HEPES pH7.5, 2M NaCl, 0.8 mM EDTA), overlaid with 10 µl of mineral oil, and incubated at 78 °C for 22 h. Five µl of pre-warmed DSN buffer (0.1 M Tris pH 8.0, 10 mM $MgCl_2$, 2 mM DTT) were then added to each sample. After a five-minute incubation at 70 °C, 0.1 U of DSN enzyme was added to the samples, and they were incubated for a further 15 min at 70 °C. Digestion was stopped by adding 10 µl of 10 mM EDTA. The fragments were then further amplified using Kapa HiFi HotStart ReadyMix (Roche) and primers that completed the adapters and added a six-base pair index to the P7 adaptor. All adaptor and primer sequences are reported in Supplementary Table 3. After amplification, the libraries were purified with 1 volume of SPRI beads, quantified using a QuBit dsDNA Broad Range Assay Kit (Invitrogen) and analysed on a 2100 Bioanalyzer instrument using a High Sensitivity DNA Analysis Kit (Agilent).

All libraries were sequenced at the McGill University and Génome Québec Innovation Center on HiSeq2500, HiSeq4000 and HiSeqX instruments (Illumina), to produce paired-end, 150-bp reads. Libraries with fewer reads were resequenced to increase genome coverage. After quality filtering, a total of 60.7 billion read pairs were retained, equivalent to 14.5 Tbp of sequence data.

Variant calling

Variants were called on a set of individuals that included the 1,518 samples described in 'DNA isolation, library preparation and sequencing', a set of cultivated *H. annuus* lines[16] and wild *Helianthus* samples previously sequenced for other projects[16,58,59], for a total of 2,392 samples (Supplementary Table 1). The additional samples were included to improve SNP calling, and to identify haploblock genotypes. Sequences were trimmed for low quality using Trimmomatic[60] (v0.36) and aligned to the *H. annuus* XRQv1 genome[9] (HanXRQr1.0-20151230) using NextGenMap[61] (v.0.5.3). The resulting SAM files were converted to BAM, concatenated and sorted (samtools[62,63] v.0.1.19); PCR duplicates were marked (picard[64] MarkDuplicates 2.9.3) and the BAM file was indexed. For libraries sequenced in multiple lanes, BAM files were merged by sample identifier (sambamba[65] v,0.6.6) and PCR duplicates were remarked.

To perform variant calling, we followed the best practices recommendations of the Genome Analysis ToolKit (GATK)[66], and executed steps documented in GATK's germline short variant discovery pipeline (for GATK 4.0.1.2). To reduce computational time and improve variant quality, we excluded genomic regions containing transposable elements[9], which represent about three-quarters of the sunflower genome, and to which short reads cannot be reliably mapped. The callable regions comprised 1.1 Gbp of the total 3.6 Gbp of the XRQv1 assembly[9]; the corresponding bed file is included in the code repository (HanXRQr1.0-20151230_allTEs_ubc.non-repetitive-regions.2017.sorted.bed). All downstream analyses were conducted on this transposable-element-filtered data set. HaplotypeCaller (v.4.0.1.2) was used on each sample individually to produce a genomic VCF (g.vcf). Heterozygosity settings for HaplotypeCaller step were increased to mu = 0.01 and st_dev = 0.1. This is 10-fold higher than the default, but better reflects the expected diversity in sunflowers compared to humans. HaplotypeCaller is a compute-intensive process that can take

advantage of parallelism. To speed up the HaplotypeCaller phase, the callable regions of the genome were evenly split into 160 contiguous, non-overlapping genomic intervals. For each sample, those intervals were then processed in parallel, according to the number of cores available on the compute node. The 160 resulting g.vcfs were gathered into a single per-sample g.vcf, and then indexed using tabix and bgzip (v.0.2.5-0). Joint genotyping of all samples in the same VCF would be ideal, as it allows for greater confidence on low-frequency variants and simplifies comparisons between groups of samples. An initial attempt to jointly genotype all samples for 10 random 1-Mbp windows completed; however, given the large number of samples, high levels of genetic variation and large genome size, it would have been computationally difficult to carry this operation across the genome given the available resources. Samples were therefore subdivided by species in three cohorts: *H. annuus*, *H. argophyllus* and *H. petiolaris*, which were independently genotyped. The *H. annuus* cohort included 309 cultivated and landrace *H. annuus* that were used for quality-control testing, but were removed for further analyses, before the final filtering for minor allele frequency (MAF) and missing data (see details in 'Variant quality filtering').

Before further analysis, the g.vcf files were converted into a modified TileDB format[67] using GATK's GenomicsDBImport (v.4.0.1.2). This step aggregates variants in a genomic region of interest from all samples in a cohort, and was found to be necessary to allow the next steps in the analysis to proceed. This operation was parallelized over 4-Mbp regions of the genome. TileDBs for a given region across a cohort were then converted into an unfiltered VCF using GATK's GenotypeGVCFs (v.4.0.1.2) in mode '–use-new-qual'. The new-qual mode is the default mode in newer versions of GATK (≥4.1.1.0), and was necessary to allow SNP calling to run on our compute nodes (32- or 48-core Intel Skylake, with ≤ 256 GB of RAM). Raw VCF chunks were then gathered into roughly per-chromosome files (17 files, one for each nuclear chromosome, plus one bundle file for all 'unplaced' chromosome contigs HanXRQChr00c*, chloroplast and mitochondria) using GatherVcfs (v.4.0.1.2).

Variant quality filtering

Genotyping produced VCF files featuring an extremely large number of variant sites (222 million, 78 million and 167 million SNPs and indels for *H. annuus*, *H. argophyllus* and *H. petiolaris*, respectively, combining SNPs and indels). Over the called portion of the genome, this corresponds to 0.07 to 0.2 variants per bp, with 30–47% per cent of variable sites being indel variation. The proportion of multiallelic variant sites was also notably high, varying between 24% and 51% across cohorts. To remove low-quality calls and produce a dataset of a more manageable size, we used GATK's VariantRecalibrator (v.4.0.1.2), which filters variants in the call set according to a machine-learning model inferred from a small set of 'true' variants. The model computed by the recalibrator attempts to define boundaries in the multidimensional site quality space that capture all or most known variant sites. Unknown variants that fall within this boundary are included, and those outside of the boundary are removed. In this way stringency is determined by choosing the proportion of the known sites to be included in the boundary, which in GATK nomenclature is called the tranche. By selecting a smaller tranche (for example, from 99% to 90%), the model selects a more stringent boundary and produces a smaller number of more confident sites.

In the absence of an externally validated set of known sunflower variants to use as calibration, we computed a stringently filtered set from top-N samples with highest sequencing coverage for each species (n = 67 cultivated samples for *H. annuus*, and n = 20 for the other two species). In these subsets, variants were filtered using the following parameters: mapping quality >50.0, 90% sample coverage for the site, –1.0> strand odds ratio <1.0, MAF >0.25, excess heterozygosity <5.0 (for non-cultivar lines <10.0 was used), –1.0> BaseQRankSum < 1.0, depth of coverage within one standard deviation from the mean and excess het >–4.5. The resulting SNP set was then recalibrated against the set of all variants from the entire corresponding cohort, using VariantRecalibrator (v.4.0.6.0, with resource parameters 'known = false, training = true, truth = true, prior = 10.0'). To speed up processing time, and to bring memory requirements to practical levels (that is, < 250 GB), it was necessary to preprocess the large training set before calibration; we stripped genotype information columns (with MakeSitesOnlyVcf) as the genotype columns from the VCF are not consulted by VariantRecalibrator. Following recommended practices, an early filtering pass to remove sites with extremely unlikely heterozygosity (excess het z-score <4.5) was also performed.

The stringency of the algorithm in classifying true or false variants was adjusted by comparing variant sets produced for different parameter values (tranche 100.0, 99.0, 90.0, 70.0 and 50.0). For each cohort, results for tranche = 90.0 were chosen for downstream analysis, based on heuristics: the number of novel SNPs identified, and improvements to the transition/transversion ratio (towards GATK's default target of 2.15). Filtering by tranche retained 13.1%, 24.5% and 30.7% of the total raw SNPs for *H. annuus*, *H. petiolaris and H. argophyllus*, respectively. The SNP data for the three species were then divided in the smaller sets used for the different analyses (GWA, GEA and so on), and filtered for MAF ≥ 0.01, genotype rate ≥ 90% and to keep only bi-allelic SNPs. The samples included in subsets used for different analyses are listed in Supplementary Table 1. Each set included a mean of 25–39 variants per gene in the genotyped regions of the genome; additional information on the SNP distribution within genic regions is reported in Supplementary Table 1.

The pipeline described in this section, including its data and software dependencies, were programmed into a Snakemake[68] (v.4.7.0) workflow. To ensure reproducibility, the pipeline also makes extensive use of conda package environments, and Docker containers with precise versioning. Calling and filtering was computed on Compute-Canada's High-Performance-Computing (HPC) Cedar cluster.

Assessing variant quality

To assess genotype accuracy, we selected 12 individuals from each species (one randomly chosen individual from each of 12 populations spanning the whole range of each species), PCR-amplified six approximately 1-kbp regions from the same DNA that was used for library construction, and determined their sequence by Sanger sequencing. We then compared our next-generation-sequencing-based genotypes to the Sanger sequencing results, and determined the percentage of genotype matches at different sequencing depths in our VCF file (Extended Data Fig. 1d, Supplementary Table 1). For *H. petiolaris*, six individuals each from subspecies *petiolaris* and subspecies *fallax* were selected. Primers for PCRs were designed in exons, to maximize the chances that the PCRs would be successful across all the individuals for a species, and PCR products spanned at least one intron. All PCR and sequencing primers are reported in Supplementary Table 3.

Remapping sites to the HA412-HOv2 reference genome

Our initial analysis of haploblocks (see 'Population genomic detection of haploblocks'), as well as GWA and GEA results for haploblocks regions, found many instances of disconnected haploblocks and high linkage between distant parts of the genome, suggesting problems in contig ordering. Therefore we remapped genomic locations from XRQv1[9] to the newer HA412-HOv2 assembly[11]; to do so, 200 bp of reference sequence flanking each site in XRQv1 was extracted and aligned to HA412-HOv2 using BWA (v.0.7.17)[69]. These alignments were filtered for mapping quality >40, and the HA412-HOv2 position for the variant site was extracted. Because all remapped sites were not in repetitive regions and had passed variant quality score recalibration filtering, remapping success rate was high (96–98%). Whenever mapping suggested two different variants on the XRQv1 genome were in the same position on the HA412-HOv2 genome, probably owing to indels and imprecise alignment, one site was shifted by one bp so they did not overlap. Remapping was preferred to de novo read alignment and variant

calling against the HA412-HOv2 assembly because of the prohibitive amount of computational time that would have required. Measures of linkage disequilibrium (LD) between all sites within 200 kb on chromosome 2 using vcftools (v.0.1.13)[70] showed that remapping significantly improved LD decay (Extended Data Fig. 1a) and produced more contiguous haploblocks (Extended Data Fig. 1b), supporting the accuracy of the new genome assembly and our remapping procedure. SNPs remapped to HA412-HOv2 were therefore used for all analyses presented in this paper. Although we recognize that this approach reduces accuracy at the local scale, and would not be appropriate—for example—for determining the effects of variants on coding sequences, it produces a more accurate reflection of the genome and linkage structure.

Phylogenetic analysis

To determine phylogenetic relationships between samples, variants were called for 20 windows of 1 Mbp, randomly selected across the genome. Indels were removed and SNP sites were filtered for <20% missing data and MAF >0.1%. All sites were then concatenated and analysed using IQ-tree[71–73] with ascertainment bias and otherwise default parameters. On the basis of the results of the phylogenetic analysis, cases in which samples grouped outside their assumed population or species were reassigned if a source of error was confidently identified (that is, mislabelling during DNA extraction, library preparation or sequence analysis). Otherwise, the sample was removed. Samples with more intermediate phylogenetic positions were not removed, as they could represent admixed ancestry rather than misidentification.

Genome-wide association mapping

Samples that were sequenced but were not part of the common garden experiment were removed from the variants dataset before filtering for MAF ≥ 3%. Variants were imputed and phased using Beagle[74] (version 10Jun18.811). Genome-wide association analyses were performed for 86, 30 and 69 phenotypic traits in *H. annuus*, *H. argophyllus* and *H. petiolaris*, respectively, using the EMMAX (v.07Mar2010) or the EMMAX module in EasyGWAS (v.2.9)[75]; both approaches use the same method and produced comparable results. Population structure was controlled for by including the first three principal components as covariates, as well as an IBS kinship matrix calculated by EMMAX[76]. For every SNP or peak above the Bonferroni significance threshold, genes within a 100-kbp interval centred in the SNP with the lowest *P* value, or within the boundaries of the GWA peak (whichever is larger), are reported in Supplementary Table 2. Inflorescence and seed traits could not be collected for *H. argophyllus*, because most plants of this species flowered very late in our common garden, and did not form fully developed inflorescences and set seeds before temperatures became too low for their survival.

Genome–environment association analyses

Twenty-four topoclimatic factors were extracted from climate data collected over a 30-year period (1961–1990) for the geographical coordinates of the population collection sites, using the software package Climate NA[77]. Soil samples from each population were also analysed for 15 soil properties (Supplementary Table 1). The effects of each environmental variable were analysed using BayPass[78] version 2.1. Population structure was estimated by choosing 10,000 putatively neutral random SNPs under the BayPass core model[78]. The Bayes factor (denoted BF_{is} as in ref. [78]) was then calculated under the standard covariate model to evaluate the association of SNP frequencies with 39 geographical, climatic and soil variables. For each SNP, BF_{is} was expressed in deciban units (dB $10 \log_{10}(BF_{is})$). Population PET_30 was removed from GEA analyses of *H. petiolaris petiolaris*, as very divergent haplotypes on two chromosomes made it an extreme outlier in the population correlation matrix, which resulted in GEA association values that were overall much lower than in the other three datasets. Populations ANN_71 and PET_21 were removed from the soil GEA analyses because no soil samples were available for them.

To calculate a significance threshold for candidate gene identification, pseudo-observed data (POD) were used with the random 10,000 SNPs used for the core model, and a 1% empirical threshold was calculated for the observed Bayes factor. This value ranged from 6.7 to 7.3 depending on the species, and produced an extremely large number of outlier regions. We therefore followed ref. [78] and used Jeffreys' rule[79], quantifying the strength of associations between SNPs and variables as 'strong' ($10 \text{ dB} \le BF_{is} < 15 \text{ dB}$), 'very strong' ($15 \text{ dB} \le BF_{is} < 20 \text{ dB}$) and 'decisive' ($BF_{is} \ge 20 \text{ dB}$). To produce a narrower set of candidate genes, the top 10 non-overlapping 50 SNP windows based on the median BF_{is} value were selected for each species and variable. A list of all the genes within these windows with at least one SNP with $BF_{is} \ge 20$ dB within 1 kbp of their boundaries is reported in Supplementary Table 2.

Transgenes and expression assays

Total RNA was isolated from mature leaves and apical meristems using TRIzol (Thermo Fisher Scientific) and complementary DNA (cDNA) was synthesized using the RevertAid First Strand cDNA Synthesis kit (Thermo Fisher Scientific). The complete coding sequences (CDS) of Ha*FT1*, Ha*FT2* and Ha*FT6* were amplified from cDNA from *H. argophyllus* individuals carrying the early and late haplotype for arg06.01. Two alleles of the Ha*FT2* CDS were identified in late-flowering *H. argophyllus* plants (one of them identical to the Ha*FT2* CDS from early-flowering individuals), differing only for two synonymous substitutions at position 285 and 288. All alleles were placed under control of the constitutive CaMV 35S promoter in pFK210 derived from pGREEN[80]. Constructs were introduced into plants by *Agrobacterium-tumefaciens*-mediated transformation[81]. Col-0 and *ft-10* seeds were obtained from the *Arabidopsis* Biological Resource Center. All primer sequences are reported in Supplementary Table 3.

Population genomic detection of haploblocks

The program lostruct (local PCA/population structure, v.0.0.0.9) was used to detect genomic regions with abnormal population structure[29]. Lostruct divides the genome into non-overlapping windows and calculates a PCA for each window. It then compares the PCAs derived from each window and calculates a similarity score. The matrix of similarity scores is then visualized using a multidimensional scaling (MDS) transformation. Lostruct analyses were performed on the *H. annuus*, *H. argophyllus*, *H. petiolaris petiolaris* and *H. petiolaris fallax* datasets, as well as in a *H. petiolaris* dataset including both *H. petiolaris petiolaris* and *H. petiolaris fallax* individuals. For each dataset, lostruct was run with 100 SNP-wide windows and independently for each chromosome. Each MDS axis was then visualized by plotting the MDS score against the position of each window in the chromosome.

Many localized regions of extreme MDS values with high variation in MDS scores and sharp boundaries were detected (Fig. 4a, Extended Data Fig. 4). Localized changes to population structure could occur owing to selection or introgression, but both the size and discrete nature of the regions are consistent with underlying structural changes defining the boundaries and preventing recombination. For example, inversions prevent recombination between orientations and if inversion haplotypes are diverged enough, they will show up in lostruct scans[29]. Because we are interested in recombination suppression in the context of adaptation, we focused on regions that had the following features: (1) a PCA in the region should divide samples into three groups representing 0/0, 0/1 and 1/1 genotypes, (2) the middle 0/1 genotype should have higher average heterozygosity and (3) there should be high LD within the region. We focused on the regions that best fit this expectation by manually curating the list of regions. Other processes, such as linked selection, can produce inversion-like patterns in the lostruct output so we were unable to automate inversion discovery.

Potential haploblock regions were defined on the basis of MDS plots, and an MDS axis and minimum or maximum value that included windows within the region, but excluded the rest of the chromosome,

were manually selected. Because there was variation in MDS score within each region, and an individual window within the region may fall below the cut off, windows that were surrounded by selected windows, within a range of 20 windows, were included. In most cases this resulted in a single unbroken range, but some regions—mainly *H. argophyllus* and *H. petiolaris*—were broken into multiple nearly abutting ranges. Furthermore, for *H. petiolaris* several of the regions were broken into unconnected distant regions, which probably reflects rearrangements in the *H. petiolaris* genome relative to the *H. annuus* reference used (Extended Data Fig. 6c).

All SNPs within the regions defined by MDS scores were used to calculate PCAs using SNPrelate[82]. The *k*-means clustering algorithm in R was used to define three clusters from PC1[83,84]. Because sample sizes were often unbalanced between the three potential groups, the starting positions for the three clusters were chosen as the maximum, minimum and middle of the range of PC1 scores. *K*-means cluster assignment was used as a preliminary genotype for the sample. Observed heterozygosity was also measured in each group. For all retained regions, samples clearly fell into three groups and observed heterozygosity was higher in the middle (0/1) group.

To visualize LD patterns, all SNPs with MAF <5% were removed, the remaining variants were thinned to one per 100 bp, and genotype R^2 values for all sites within a chromosome were calculated. Values were grouped into 500-kbp windows and the second largest R^2 value was plotted (Fig. 4e, Extended Data Fig. 4). In each case, regions identified in lostruct had high LD.

The combined evidence of PCA and linkage suggests that the lostruct outlier regions are characterized by long haplotypes with little or no recombination between haplotypes. We refer to these as haploblocks. To explore the haplotype structure underlying the haploblocks, sites correlated ($R^2 > 0.8$) with principal component 1 (PC1) in the PCA of the haploblock were extracted as haplotype diagnostic sites and used to genotype the haploblocks. Because there is seemingly little recombination between haplotypes, this is conceptually similar to a hybrid index and we expect all samples to be consistently homozygous for one haplotype's alleles or be heterozygous at all sites (that is, similar to an F_1 hybrid). Haploblock genotypes were assigned to all samples using $0/0 = p < 0.5$, $h \leq (-2/3)p + (2/3)$; $1/1 = p \geq 0.5$, $h \leq (2/3)p$; else 0/1, in which p is the proportion of haplotype 1 alleles and h is the observed heterozygosity. The haplotype structure was also visualized by plotting diagnostic SNP genotypes for each sample, with samples ordered by the proportion of alleles from haplotype 1 (Fig. 2f).

The underlying recombination landscape in haploblock regions was explored by subsetting our dataset to samples homozygous for the more common haploblock genotype and measuring LD across the region. As before, SNPs with MAF <5% were removed, variants were thinned to one per 100 bp and genotype R^2 values for all sites within a chromosome were calculated. If the signal of high LD is only present when both haploblock genotypes are included, then it supports mechanisms that specifically prevent recombination between haplotypes. That being said, some haploblocks fall in generally low recombination regions and high LD within a haploblock genotype does not preclude recombination suppression.

Lostruct was run on individual SNP datasets containing *H. petiolaris* subsp. *petiolaris* or *H. petiolaris* subsp. *fallax*, and both subspecies together. Although each dataset produced a collection of haploblocks, they were not identical. Some haploblocks were identified in one subspecies but not the other, and some were only identified when both subspecies were analysed together. In some cases, it was clear that haploblocks identified in both subspecies represented the same underlying haploblock because they physically overlapped and had overlapping diagnostic markers. We manually curated the list of haploblocks and merged those found in multiple datasets. We set the boundaries of these merged haploblocks to be inclusive (that is, include windows found in either) and the diagnostic markers to be exclusive (that is, only include sites found in both). For this merged set of haploblocks, all *H. petiolaris* samples were genotyped using diagnostic markers.

Design of genetic markers for haploblock screening

Diagnostic SNPs for haploblocks were extracted from filtered VCF files. The resulting cleaved-amplified polymorphic sequence (CAPS) markers or direct sequencing markers were tested on representative subsets of individuals included in the original local PCA analysis (Fig. 4a, Extended Data Fig. 4), for which the genotype at haploblocks of interest was known. Marker information is reported in Supplementary Table 3.

Sequencing coverage analysis

To detect the presence of potential deletions in the late-flowering allele of arg06.01, SNPs in the haploblock region with average coverage of at least four across at least one of the genotypic classes were selected (to exclude positions with overall low mapping quality). SNP positions with extremely high average coverage (>15) were removed, as they are likely to represent duplicated or paralogous regions. For the analyses reported in Extended Data Fig. 2c, SNP positions with coverage 0 or 1 were considered missing data.

Comparisons of *H. annuus* reference assemblies

Masked reference sequences for the *H. annuus* cultivars HA412-HOv2 and PSC8[11,12] were aligned using MUMmer[85] (v.4.0.0b2). The programs nucmer (parameters -b 1000 -c 200 -g 500) and dnadiff within the MUMmer package were used. Only orthologous chromosomes were aligned together because of the high similarity and known conservation of chromosome structure. The one-to-one output file was then visualized in R and only included alignments in which both sequences were >5,000 bp. Inversion boundaries and sequence identity between haplotypes were further determined using Syri (v.1.0)[86].

Comparisons of genetic maps

Fourteen genetic maps were used: the seven *H. annuus* genetic maps used in the creation of the XRQv1 genome[9]; three newly generated *H. annuus* maps obtained from wild × cultivar F_2 populations (E.B.M.D., M.T., G.L.O. and L.H.R., manuscript in preparation); two previously published *H. petiolaris* genetic maps obtained from F_1 crosses[87]; and two newly generated *H. petiolaris* maps[88]. Whenever necessary, marker positions relative to XRQv1 were re-mapped to the HA12-HOv2 assembly.

Six of the previously described *H. annuus* maps were obtained from crosses between cultivars (the seventh one was obtained from a wild × cultivar cross); to determine which haploblock could be expected to segregate in the genetic maps, all of the cultivated sunflower lines were genotyped for each *H. annuus* haploblock using diagnostic markers identified in wild *H. annuus*. Ann01.01 and ann05.01 were found to be highly polymorphic among cultivated lines, and other haploblocks were fixed or nearly fixed for a single allele. For all 14 maps, marker order was compared to physical positions in the HA412-HOv2 reference assembly, and evidence for suppressed recombination or structural variation was recorded (Extended Data Table 1).

Hi-C

On the basis of our resequencing data, a pair of *H. petiolaris* and a pair *H. argophyllus* populations were selected that diverged for the largest number of haploblocks (PET_47 and PET_08 for *H. petiolaris* and ARG_18 and ARG_23 for *H. argophyllus*). Several individuals from each population were grown and genotyped at diagnostic SNPs for several haploblocks (pet09.01, pet10.01, pet10.01 and pet14.01 for *H. petiolaris*; arg06.01 and arg10.01 for *H. argophyllus*; see 'Design of genetic markers for haploblock screening') to identify, for each species, a pair of individuals with different genotypes at the largest possible number of haploblocks. Chromosome conformation capture sequencing (Hi-C)[38,89] was then performed on one individual each from these four populations, to compare the structural organization of the different

haplotypes at haploblock regions. Additionally, three Hi-C libraries from *H. annuus* HA412-HO were included in the analysis; data from these libraries were used to assemble the current HA412-HOv2 reference genome[11], and are used here as an interaction baseline. All libraries were prepared by Dovetail Genomics, and each library was sequenced on a single lane of HiSeq X with 150-bp paired-end reads. Given the size and repetitive nature of sunflower genomes, Hi-C data could not be used to assemble a full genome for the wild sunflower samples; the HA412-HOv2 cultivated sunflower assembly was therefore used as a reference, and patterns of interactions were compared between samples. Reads were trimmed for enzyme cut site (DpnII) and base quality using the tool trim in the package HOMER[90] (v.4.10) with the following flags: '-3 GATC -mis 0 -matchStart 20 -min 20 -q 15'. Trimmed data were then aligned to the HA412-HOv2 reference genome using NextGenMap[61] (v.0.5.4) and interactions were quantified using the calls 'makeTagDirectory -tbp 1 -mapq 10' and 'analyzeHiC -res 1000000 -coverageNorm' from HOMER. This removes PCR duplicates on the basis of mapping location, requires reads to have ≥ 10 mapping quality and normalizes interactions in 1-Mbp windows based on the total number of interactions. To determine which haploblocks differ between samples, aligned sequence data and samtools mpileup[63] were used to genotype diagnostic markers and call genotype for each haploblock, as described in 'Population genomic detection of haploblocks'.

Hi-C data were used in two ways to identify structural changes. First, the difference between interaction matrices for samples of the same species was plotted for each haploblock region where the two samples had different genotypes. Second, the difference between interaction matrices for *H. annuus* (using the HiC data that were generated to scaffold the HA412-hOv2 reference assembly[11]) and each *H. petiolaris* and *H. argophyllus* sample were plotted. We identified and highlighted long-distance interactions that differed between samples and that were consistent with structural variations underlying haploblocks. To determine how common these interactions are, we compared the difference in interaction strength at the identified windows with all windows of the same genomic distance.

Interpretation of the HiC patterns was sometimes complicated by the presence of putative structural differences between the genome of *H. petiolaris* and that of the HA412-HOv2 reference assembly against which reads from the HiC libraries were mapped. To determine what HiC patterns would be expected in those situations if haploblocks are associated with large inversions, we simulated an interaction matrix in which interactions between windows linearly decayed on the basis of distance. We then flipped window ordering within a region to simulate an inversion, and compared the interaction matrices with the original and flipped ordering. We used these basic HiC simulations to produce possible rearrangements between the haploblocks in *H. petiolaris* and the *H. annuus* reference that fit the observed HiC interaction patterns for three representative haploblocks (Extended Data Fig. 7b).

Haploblock phenotype and environment associations

Because haploblocks are large enough to affect genome wide population structure, their associations with phenotypes of environmental variables may be masked when controlling for population structure. Therefore, a version of the variant file was created with all haploblock sites removed; both sites within haploblock regions and sites in close linkage (vcftools[70] v.0.1.14, $R^2 > 0.5$) with haploblock genotypes were removed, to make sure that sites that were physically within the haploblock region were removed even if they were placed elsewhere owing to reference differences. This haploblock-removed version of the genotype file was used for calculating PCA and kinship for EMMAX and the genetic covariance matrix for BayPass.

GWA analyses were performed using EMMAX[76] (v.07Mar2010) for all traits measured in the common garden experiment (Supplementary Table 1). For all runs, the first three PCs were included as covariates, as well as a kinship matrix calculated from the haploblock-removed genotype table. Environmental associations were run using BayPass[78] as previously described (see 'Genome-environment association mapping'), except that the 10,000 SNPs used to estimate population structure were drawn from the haploblock-removed dataset. Regions of high associations colocalizing with haploblock regions were identified, and haploblocks were also directly tested by coding each haploblock as a single biallelic locus.

To examine the relative importance of haploblocks to trait evolution and environmental adaptation, association results were compared between haploblocks and SNPs. Using SNPs as a baseline allows controlling for the correlation between traits or environmental variables. To make values comparable, both SNPs and structural variants (SVs) with MAF ≤ 0.03 were removed. Each locus was classified as associated ($P < 0.001$ or $BF_{is} > 10$ dB) or not to each trait. The number of traits or climate variables each locus was associated with was then counted. The proportion of loci with ≥ 1 traits/climate variables associated for SNPs and haploblocks was then compared using prop.test in R[84] (Extended Data Fig. 9b).

Haploblock phylogenies and dating

A phylogenetic approach was used to determine the divergence time between haploblock alleles. For each haploblock, five samples homozygous for each haploblock allele were chosen (defined as having >85% SNP ancestry from one haploblock allele). Two random samples from the other (sub)species, as well as two perennial samples (*H. grosseserratus* and *H. divaricatus*) were included in the analyses. For *H. petiolaris*, subsp. *petiolaris* and subsp. *fallax* were included in the same phylogeny if a haploblock was segregating in both. All genes within the haploblock in the HA412-HOv2 genome annotation were extracted, and the corresponding gene regions in the XRQv1 assembly were identified using a list of one-to-one orthologues between the two assemblies, created using Swiftortho[91]. For each gene, gVCF files were created from BAM files of the samples with GATK's (v.4.0.6.0) HaplotypeCaller and gene sequences in FASTA format were generated using a custom Perl script. Haploblocks with more than 100 genes were down-sampled to 100 genes to reduce computing time.

The phylogeny of each haploblock region was estimated by Bayesian inference using BEAST[92] 1.10.4. The dataset was partitioned, assuming unlinked substitution and clock models for the genes, and analysed under the HKY model with 4 gamma categories for site heterogeneity: a strict clock, a 'constant size' tree prior with a gamma distribution with shape parameter 10.0 and a scale parameter 0.004 for the population size. Default priors were used for the other parameters. A custom Perl script was used to combine FASTA sequences and the model parameters into XML format for BEAST input. The Markov chain Monte Carlo process was run for 1 million iterations and sampled every 1,000 states. The convergence of chains was inspected in Tracer[93] 1.7.1. To estimate divergence times, the resulting trees were calibrated using a mutation rate estimate of 6.9×10^{-9} substitutions per site per year for sunflowers[94], and visualized with R package ggtree[95] and Figtree v.1.4.4[96]. Divergence times were extracted from the trees and plotted showing the 95% highest posterior density interval based on the BEAST posterior distribution. This was repeated for 100 nonhaploblock genes to estimate the species divergence times.

For the 10-Mbp region on chromosome 6 controlling flowering time in *H. argophyllus*, the early-flowering haplotype grouped with *H. annuus*. To determine whether it is the product of an ancient haplotype that has retained polymorphism only in *H. annuus* or whether it is introgressed from *H. annuus*, the phylogeny of 10 representative *H. argophyllus* samples homozygous for each haploblock allele, as well as 200 *H. annuus* samples, was inferred using IQ-tree (v.1.6.10). SNPs from the 10-Mbp region were concatenated and the maximum likelihood tree was constructed using the GTR model with ascertainment bias correction. Branch support was estimated using ultrafast bootstrap implemented in IQ-tree[71–73] with 1,000 bootstrap replicates.

Phylogenies of haploblock arg03.01, arg03.02 and arg06.02 were inferred using the same approach. To explore intraspecific history of the *H. petiolaris* haploblocks, all samples homozygous for either allele for each haploblock were selected, and phylogenies were constructed using IQ-tree with the same settings.

Statistical and reproducibility information for Figs. 1–5

Figure 1. In Fig. 1c, days to bud GWAs were calculated using two-sided mixed models. $n = 612$ individuals. Only positions with $-\log_{10} P$ value >2 are plotted. In Fig. 1d, DD <18 GEAs were calculated using two-sided XtX statistics. $n = 71$ populations. Only positions with $BF_{is} > 9$ dB are plotted.

Figure 2. In Fig. 2c, d, days to bud GWAs were calculated using two-sided mixed models, and dominant allele encoding. $n = 277$ individuals. Only positions with $-\log_{10} P$ value >2 are plotted. In Fig. 2e, number of individuals: $n = 265$ (cluster 0); $n = 27$ (cluster 1); $n = 7$ (cluster 3). In Fig. 2g, number of individuals: $n = 242$ (0/0); $n = 25$ (0/1); $n = 11$ (1/1); $n = 586$ (*ann* = *H. annuus*). Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. In Fig. 2h, number of individuals: $n = 261$ (0/0); $n = 25$ (0/1); $n = 12$ (1/1). Sequencing depth was comparable across haplotypes for the other four HaFT genes on chromosome 6. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. In Fig. 2i, experiments were repeated on three independent pairs of individuals, with similar results. In Fig. 2j, experiments were repeated on three independent pairs of individuals, with similar results. In Fig. 2k, similar phenotypic effects were observed across all the 32 independent *A. thaliana* Ha*FT1 ft-10* transgenic events that were generated. In Fig. 2l, number of individuals: $n = 28$ (Col-0); $n = 25$ (*ft-10*); $n = 32$ independent transgenic events (Ha*FT1 ft-10*). Statistical significance for differences in flowering time between *ft-10* and Ha*FT1 ft-10* was calculated using one-way ANOVA with post hoc Tukey HSD test, $F = 596$, df = 2. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges.

Figure 3. In Fig. 3c, number of individuals: $n = 14$ (non-dunes, common garden); $n = 10$ (dunes, common garden); $n = 57$ (non-dunes, wild-collected); $n = 53$ (dunes, wild-collected). Statistical significance for phenotypic differences was calculated using two-sided Mann–Whitney *U*-tests. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. In Fig. 3d, number of individuals: $n = 15$ (non-dunes); $n = 18$ (dunes). Statistical significance for phenotypic differences was calculated using two-sided Mann–Whitney *U*-tests. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. In Fig. 3e, seed size ($n = 165$ individuals) and flowering time ($n = 211$ individuals) GWAs were calculated using two-sided mixed models. CEC GEA were calculated using two-sided XtX statistic. $n = 23$ populations. Only positions with $BF_{is} > 9$ dB or $-\log_{10} P$ value >2 are plotted.

Figure 4. In Fig. 4c, number of individuals: $n = 272$ (cluster 0); $n = 253$ (cluster 1); $n = 388$ (cluster 2). In Fig. 4d, number of individuals: $n = 272$ (cluster 0); $n = 253$ (cluster 1); $n = 388$ (cluster 2). Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges.

Figure 5. In Fig. 5c, GWAs were calculated using two-sided mixed models and GEAs were calculated using two-sided XtX statistics. Number of individuals for GWAs: $n = 614$ (*H. annuus*); $n = 294$ (*H. argophyllus*); $n = 209$ (*H. petiolaris fallax*); $n = 163$ (*H. petiolaris petiolaris*). Number of populations for GEAs: $n = 71$ (*H. annuus*); $n = 30$ (*H. argophyllus*); $n = 23$ (*H. petiolaris fallax*); $n = 17$ (*H. petiolaris petiolaris*). In Fig. 5c, temperature difference (TD) GEAs were calculated using two-sided XtX statistics. $n = 71$ populations. Only positions with $BF_{is} > 9$ dB are plotted.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All raw sequenced data are stored in the Sequence Read Archive (SRA) under BioProject accessions PRJNA532579, PRJNA398560 and PRJNA564337. SRA accession numbers for individual samples are listed in Supplementary Table 1 (tabs 'Coverage and analyses', 'Outgroups', 'Samples from other studies' and 'HiC samples'). The HA412-HOv2 and PSC8 genome assemblies are available at https://sunflowergenome.org/ and https://heliagene.org/. Filtered SNP datasets are available at https://rieseberglab.github.io/ubc-sunflower-genome/. GWA results, as well as the corresponding SNP and trait data, are available at https://easygwas.ethz.ch/gwas/myhistory/public/20/, https://easygwas.ethz.ch/gwas/myhistory/public/21/, https://easygwas.ethz.ch/gwas/myhistory/public/22/, https://easygwas.ethz.ch/gwas/myhistory/public/23/. Ha*FT1*, Ha*FT2* and Ha*FT6* sequences have been deposited in GenBank under accession numbers MN517758–MN517761. Source data for all figures are provided at https://github.com/owensgl/haploblocks/. Source data are provided with this paper.

Code availability

All code associated with this project is available at https://github.com/owensgl/haploblocks/.

50. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
51. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
52. Rodríguez, G. R. et al. Tomato Analyzer: a useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. *J. Vis. Exp* **37**, e1856 (2010).
53. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
54. Zeng, J., Zou, Y., Bai, J. & Zheng, H. Preparation of total DNA from recalcitrant plant taxa. *Acta Bot. Sin.* **44**, 694–697 (2002).
55. Rowan, B. A., Patel, V., Weigel, D. & Schneeberger, K. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3* **5**, 385–398 (2015).
56. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
57. Matvienko, M. et al. Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride. *PLoS ONE* **8**, e55913 (2013).
58. Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L. & Rieseberg, L. H. An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* **221**, 515–526 (2019).
59. Owens, G. L., Baute, G. J., Hubner, S. & Rieseberg, L. H. Genomic sequence and copy number evolution during hybrid crop development in sunflowers. *Evol. Appl.* **12**, 54–65 (2019).
60. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
61. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
62. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
63. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
64. Broad Institute. Picard tools, http://broadinstitute.github.io/picard/ (Broad Institute, 2019).
65. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
66. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at https://www.biorxiv.org/content/10.1101/201178v3 (2017).

67. Datta, K., Gururaj, K., Naik, M., Narvaez, P. & Rutar, M. GenomicsDB: storing genome data as sparse columnar arrays. https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/genomics-storing-genome-data-paper.pdf (2017).
68. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
69. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
70. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
71. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
72. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
73. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
74. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
75. Grimm, D. G. et al. easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell* **29**, 5–19 (2017).
76. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
77. Wang, T., Hamann, A., Spittlehouse, D. & Carroll, C. Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLoS ONE* **11**, e0156720 (2016).
78. Gautier, M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**, 1555–1579 (2015).
79. Jeffreys, H. *Theory of Probability* (Clarendon, 1961).
80. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. & Mullineaux, P. M. pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation. *Plant Mol. Biol.* **42**, 819–832 (2000).
81. Weigel, D. & Glazebrook, J. *Arabidopsis: A Laboratory Manual* (CSHL, 2002).
82. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
83. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: a *K*-means clustering algorithm. *J. R. Stat. Soc. C. Appl. Stat.* **28**, 100–108 (1979).
84. R Core Team. R: A language and environment for statistical computing, https://www.R-project.org/ (2019).
85. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
86. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).
87. Ostevik, K. L., Samuk, K. & Rieseberg, L. H. Ancestral reconstruction of karyotypes reveals an exceptional rate of non-random chromosomal evolution in sunflower. *Genetics* **214**, 1031–1045 (2020).
88. Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L. & Rieseberg, L. H. Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *Mol. Ecol.*, https://doi.org/10.1111/mec.15428 (2020).
89. Marie-Nelly, H. et al. High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).
90. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
91. Hu, X. & Friedberg, I. SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. *Gigascience* **8**, giz118 (2019).
92. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
93. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
94. Sambatti, J. B., Strasburg, J. L., Ortiz-Barrientos, D., Baack, E. J. & Rieseberg, L. H. Reconciling extremely strong barriers with high levels of gene exchange in annual sunflowers. *Evolution* **66**, 1459–1473 (2012).
95. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
96. Rambaut, A. FigTree, http://tree.bio.ed.ac.uk/software/figtree/ (2009).

Acknowledgements We thank J. Gouzy and N. B. Langlade for providing access to the HA412-HOv2 annotation and PSC8 genome assembly; B. T. Moyers for discussion and providing the *H. argophyllus* picture; J. Lee-Yaw and A. J. Moreno-Geraldes for comments; D. Skonieczny, A. Kim, A. Parra and C. Konecny for assistance with fieldwork and data acquisition; A. Warfield for computing advice; J. D. Herndon for providing the dune *H. petiolaris* picture; D. G. Grimm for assistance with easyGWAS; UBC's Data Science Institute for support to J.S.L.; and Compute Canada for computing resources. Maps were realized using tiles from Stamen Design (https://stamen.com), under CC BY 3.0, from data by OpenStreetMaps contributors (https://openstreetmap.org), under ODbL. Funding was provided by Genome Canada and Genome BC (LSARP2014-223SUN), the NSF Plant Genome Program (IOS-1444522), the International Consortium for Sunflower Genomic Resources, Sofiproteol, an HFSP long-term postdoctoral fellowship to M.T. (LT000780/2013) and a Banting postdoctoral fellowship to G.L.O.

Author contributions L.H.R., S.Y., J.M.B., L.A.D. and N.B. conceived the study; D.O.B. collected seeds and soil samples from wild populations; N.B., M.T., D.O.B., I.I. and W.C. performed the common garden experiment, and collected and organized phenotypic data; L.A.D. analysed the soil samples; N.B., M.T. and E.B.M.D. generated resequencing data; M.T. and M.A.P.-R. analysed Ha*FT1* amplification and expression in *H. argophyllus*; M.T., K.H. and M.A.P.-R. generated material for HiC experiments; M.T. generated and analysed *A. thaliana* Ha*FT* transgenic lines; J.-S.L., M.N. and G.L.O. performed read alignments, and SNP calling and filtering; M.T., G.L.O., S.S., K.H., K.L.O., E.B.M.D., K.L. and M.J. analysed genomic data; S.E.S. and S.M. contributed resources; R.N. provided conceptual advice; M.T., G.L.O. and L.H.R. wrote the manuscript, with contributions from all of the authors.

Competing interests The authors declare no competing interests.

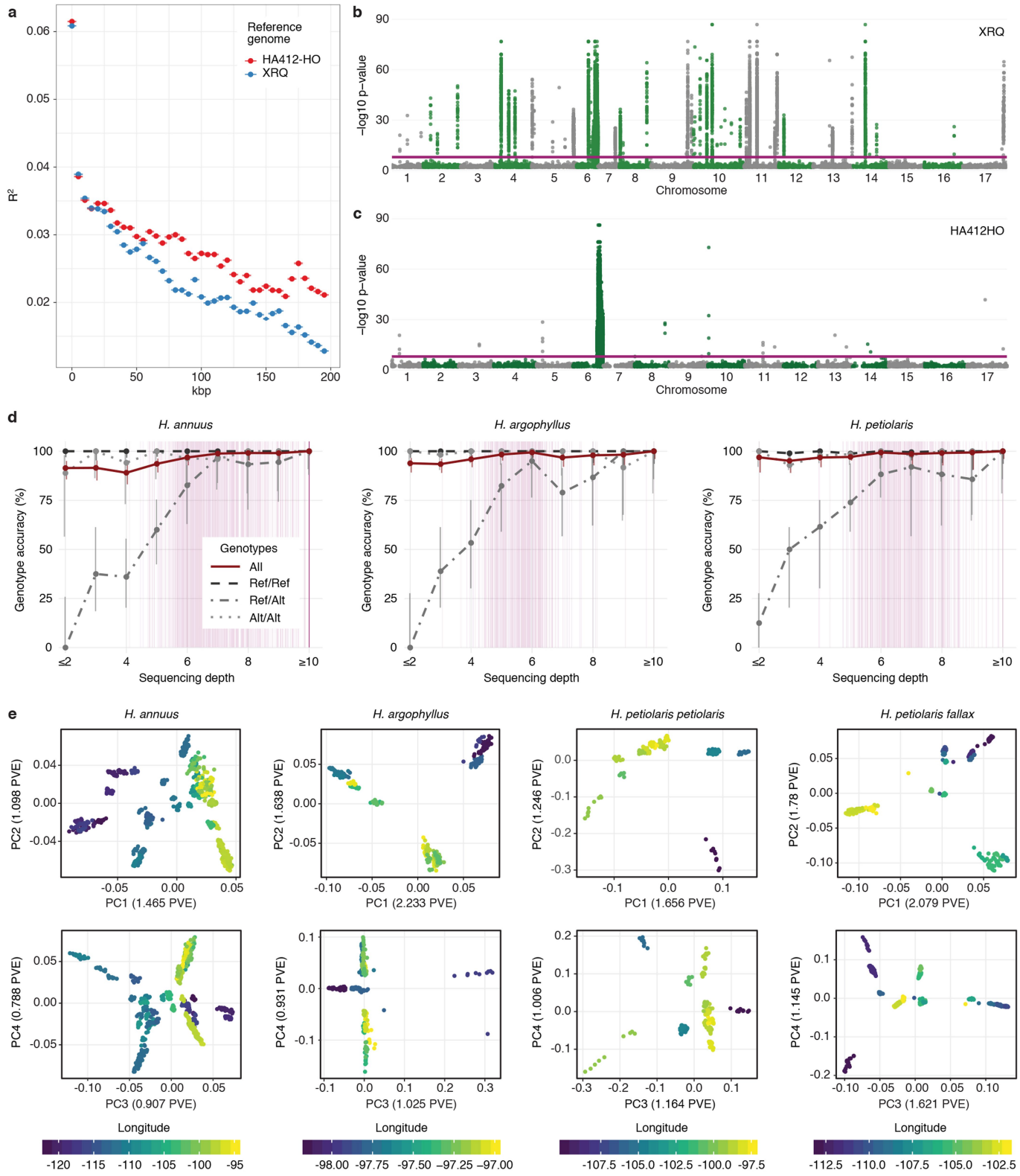
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2467-6.

Correspondence and requests for materials should be addressed to N.B. or L.H.R.

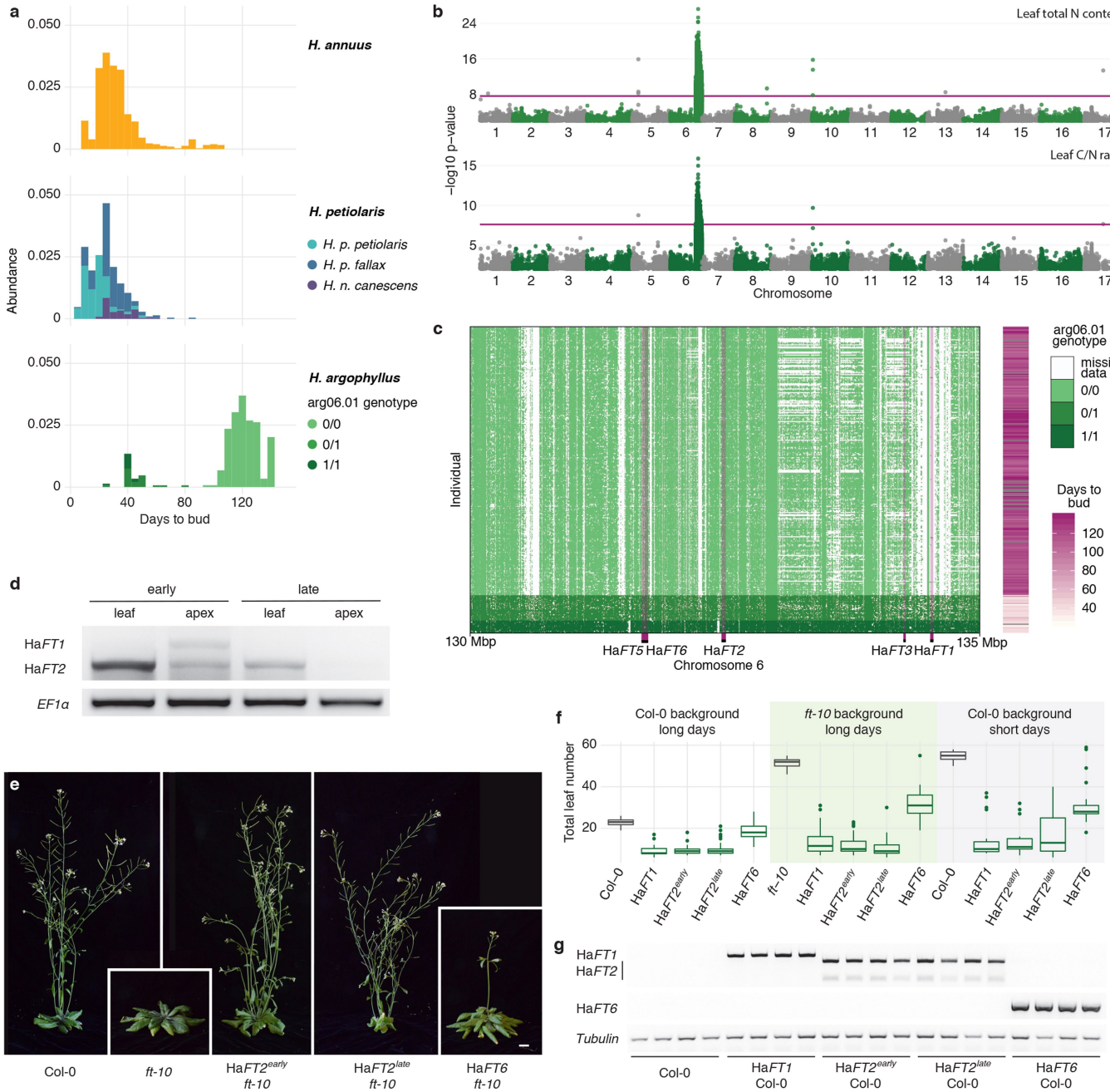
Peer review information *Nature* thanks Jeffrey Ross-Ibarra, Jeremy Schmutz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

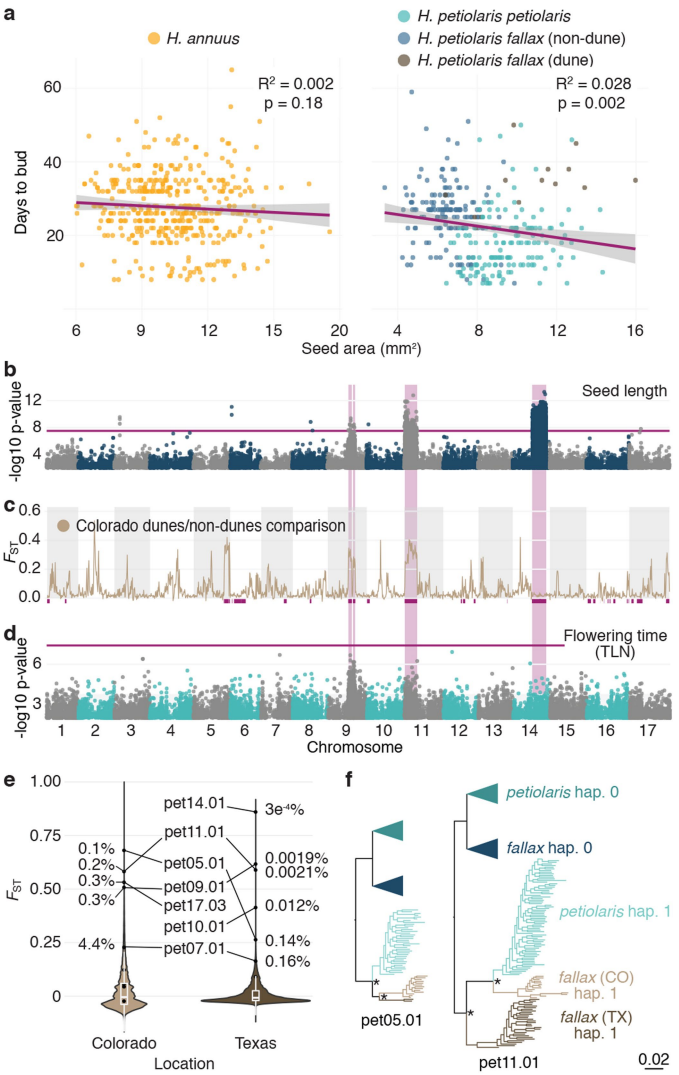


Extended Data Fig. 1 | See next page for caption.

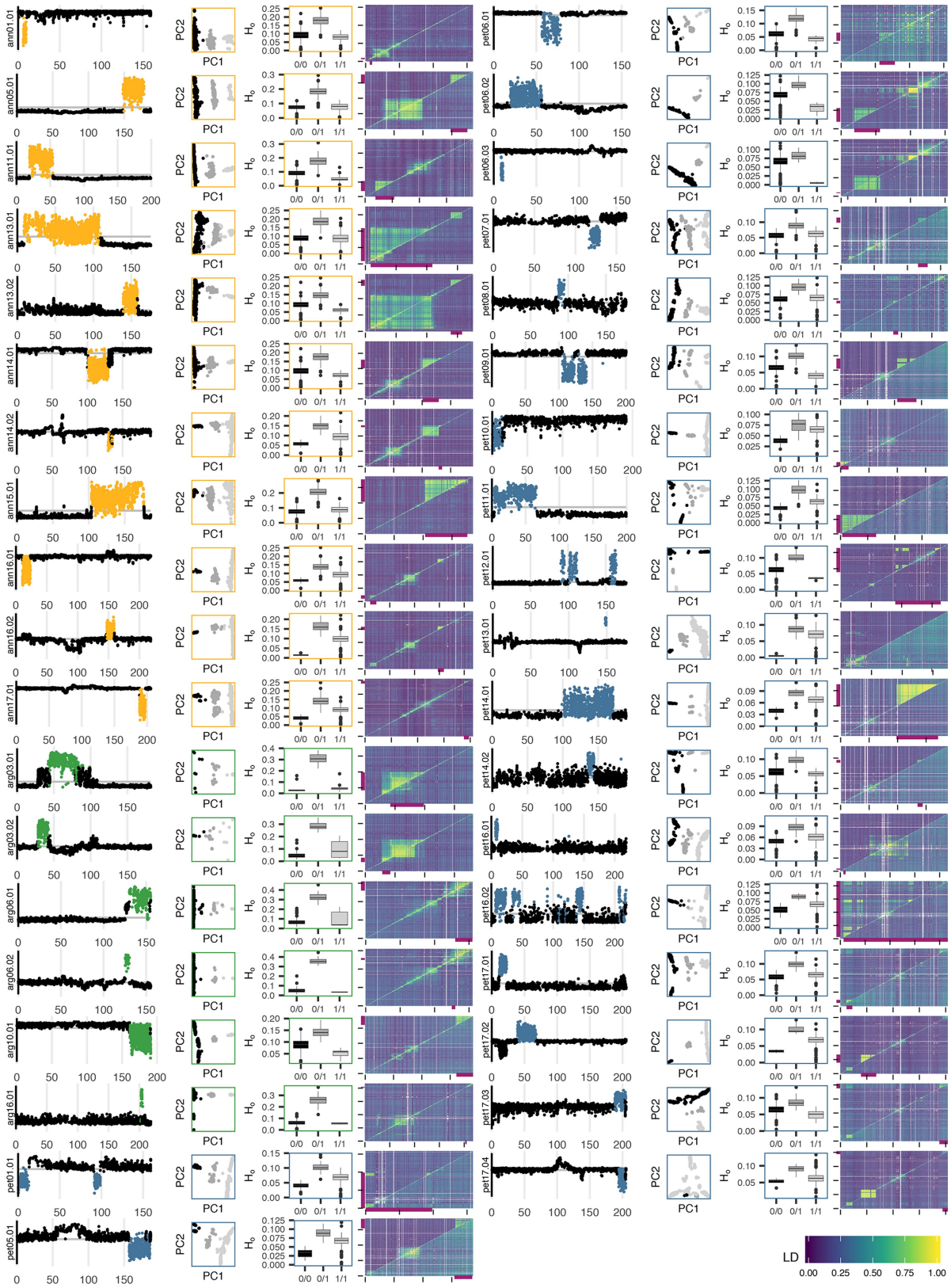
Extended Data Fig. 1 | Remapping SNPs to the HA412-HOv2 reference genome improved ordering. a, Comparison between the original order of SNPs in chromosome 2 on the XRQv1 assembly[9] (against which sequencing reads were originally mapped) and after SNP re-mapping to the HA412-HOv2 assembly[11]. Data are summarized in 5-kbp ranges. Error bars represent 2 standard errors. The higher R^2 at longer distances is due to better scaffolding of contigs in HA412-HOv2. Number of SNPs: n = 261,020 (XRQ); n 237,674 (HA412-HO). **b**, GWA for flowering time in *H. argophyllus* based on the XRQv1 assembly identified more than 40 highly significant associations. **c**, Remapping of the SNPs to the new HA412-HOv2 sunflower assembly considerably reduced the number of associations in the flowering time GWA, with the vast majority of the signal mapping to the arg06.01 haploblock region (Fig. 2). In **b**, **c**, the purple lines represent 5% Bonferroni-corrected significance. Only positions with $-\log_{10} P$ value >2 are plotted. Associations were calculated using two-sided mixed models. n = 277 individuals. **d**, Genotype call accuracy. Variants for 12 individuals for each species from our SNP dataset were compared to Sanger sequencing data. Six regions were compared. Number of sites: n = 136 (*H. annuus*); n = 139 (*H. argophyllus*); n = 262 (*H. petiolaris*). Number of genotype calls: n = 1,385 (*H. annuus*), n = 1,254 (*H. argophyllus*), n = 2,351 (*H. petiolaris*). Overall genotype accuracy: *H. annuus* = 95.9%; *H. argophyllus* = 96.8%; *H. petiolaris* = 97.9% (Supplementary Table 1). Vertical purple lines represent the average observed coverage across genic regions for individuals in the corresponding dataset. Error bars, binomial confidence interval (Wilson score method). **e**, Genome-wide principal component analysis for each dataset. Sites were pruned for linkage (r < 0.2 within 500 kb). Number of individuals: n = 730 (*H. annuus*); n = 299 (*H. argophyllus*); n = 168 (*H. petiolaris petiolaris*); n = 259 (*H. petiolaris fallax*).



Extended Data Fig. 2 | Phenotypic, structural and functional analyses for arg06.01. a, Flowering time for the three wild sunflower species measured in a common garden experiment. Number of individuals: $n = 612$ (*H. annuus*); $n = 161$ (*H. petiolaris petiolaris*); $n = 211$ (*H. petiolaris fallax*); $n = 48$ (*H. niveus canescens*); $n = 261$ (*H. argophyllus* 0/0); $n = 25$ (*H. argophyllus* 0/1); $n = 23$ (*H. argophyllus* 1/1). **b**, Leaf nitrogen content and carbon/nitrogen ratio GWAs in *H. argophyllus* (two-sided mixed model associations; $n = 289$ individuals). The purple lines represent 5% Bonferroni-corrected significance. Only positions with $-\log_{10} P$ value >2 are plotted. **c**, Genotype presence or absence for the 130–135-Mbp region of chromosome 6 in *H. argophyllus*. The *x*-axis represents consecutive SNP positions; distances on this axis are therefore not proportional to physical distances on the chromosome. Purple bars highlight the positions of the five HaFT genes in the region (Ha*FT5* and Ha*FT6* are only a few hundred bp apart). Flowering time data are the same as used in GWA analyses. **d**, Ha*FT1* and Ha*FT2* expression levels in mature leaves or shoot apices of >6-month-old, flowering *H. argophyllus* plants, grown in a greenhouse in long days conditions (14 h light:10 h dark). This experiment was performed on two independent pairs of individuals, with similar results. **e**, Six-week-old *A. thaliana* plants grown in long day conditions at 23 °C. At least 19 independent transgenic events were analysed for each construct in each genetic background, and flowering time was consistent within each group. Scale bar, 1 cm. **f**, Flowering time in long and short days (10 h light:14 h dark). Ha*FT2* alleles from early- and late-flowering *H. argophyllus* plants complement the *ft-10* mutant, similar to Ha*FT1* from the early-flowering ecotype. Ha*FT6* is expressed at low levels in *H. argophyllus* plants (not shown), and appears to be a hypo-functional *FT* homologue. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. Differences in flowering time between untransformed controls, Ha*FT6* lines and all the other transgenic lines are significant in all conditions ($P < 10^{-6}$ for all relevant comparisons; one-way ANOVA with post hoc Tukey HSD test, df = 4; exact *P* values are reported in the Source Data). Number of individuals or independent transformation events for the long days dataset in Col-0 background; $n = 28$ (Col-0); $n = 32$ (Ha*FT1*); $n = 30$ (Ha$FT2^{early}$); $n = 34$ (Ha$FT2^{late}$); $n = 45$ (Ha*FT6*). For the long days dataset in *ft-10* background: $n = 25$ (*ft-10*); $n = 30$ (Ha*FT1*); $n = 38$ (Ha$FT2^{early}$); $n = 45$ (Ha$FT2^{late}$); $n = 18$ (Ha*FT6*). For the short days dataset; $n = 10$ (Col-0); $n = 24$ (Ha*FT1*); $n = 17$ (Ha$FT2^{early}$); $n = 31$ (Ha$FT2^{late}$); $n = 31$ (Ha*FT6*). **g**, PCR detection of transgene expression in leaves of plants grown for four weeks in long days. The reduced ability of Ha*FT6* to induce flowering is not due to inefficient expression of the transgene. Results for four independent primary transformants for each transgenic line and for wild-type Col-0 plants are shown. For gel source data, see Supplementary Fig. 1.

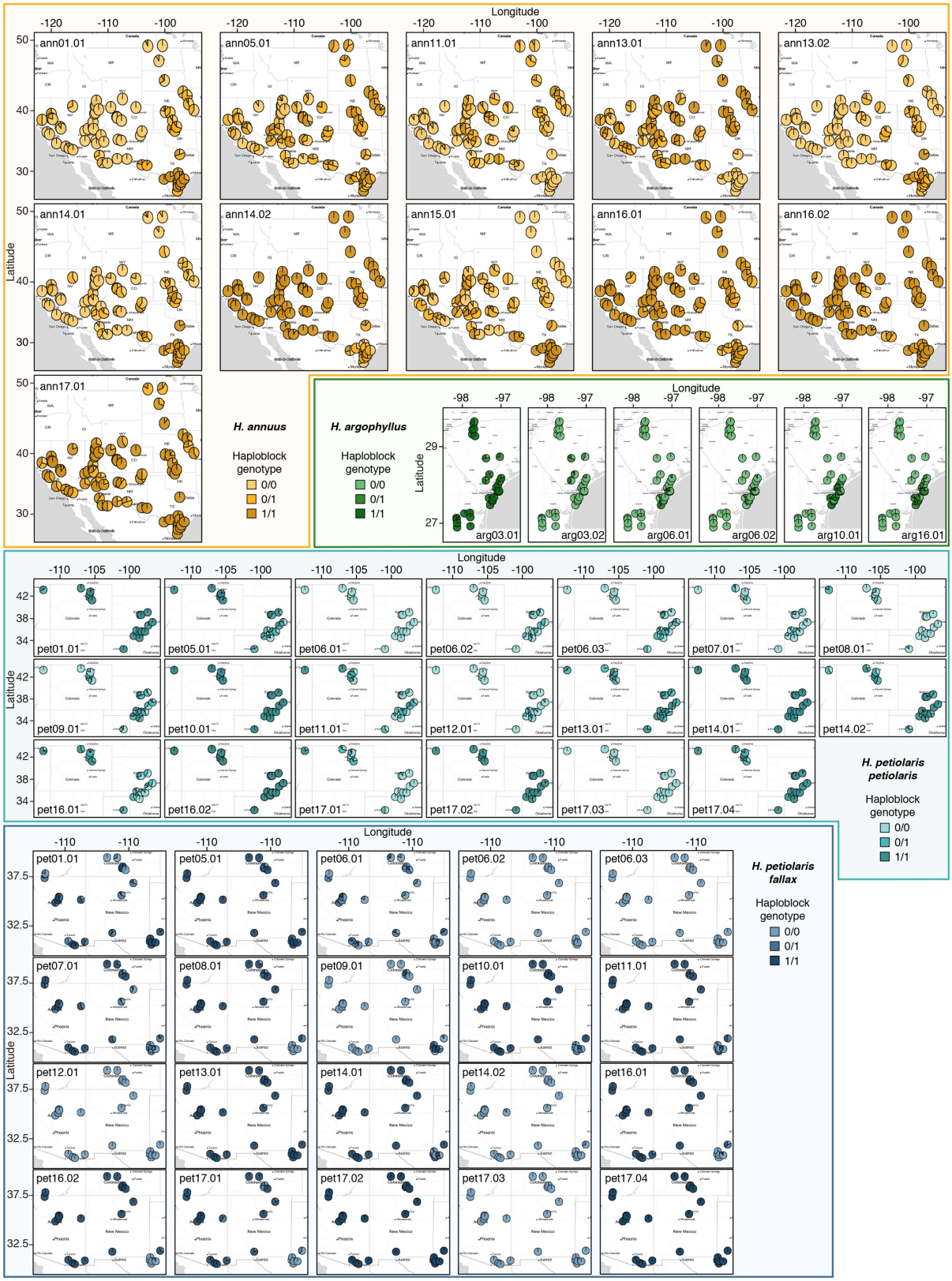


Extended Data Fig. 3 | Several haploblocks differentiate dune and non-dune populations of *H. petiolaris*. **a**, Correlation between seed size and flowering time. Although dune-adapted *H. petiolaris fallax* flowers later and has larger seeds than non-dune-adapted populations, these two traits generally show no correlation, or a weak negative correlation, in *H. annuus* and *H. petiolaris*. Purple lines represent linear regressions, shaded grey area are 95% confidence intervals. *H. annuus*: $n = 426$ individuals, one-sided $F_{1,423} = 1.831$, $P = 0.18$; *H. petiolaris*: $n = 307$ individuals, one-sided $F_{1,305} = 9.841$, $P = 0.0019$. **b**, Seed length GWA in *H. petiolaris fallax* (two-sided mixed model associations; $n = 165$ individuals). No significant association with haploblocks is found in GWA analyses for seed width (not shown). **c**, F_{ST} values in 2-Mbp non-overlapping sliding windows for comparisons between dune- and non-dune-adapted populations of *H. petiolaris fallax* in Colorado. Purple bars represent predicted haploblocks. **d**, Flowering time (approximated as total leaf number (TLN) on the primary stem) GWA for *H. petiolaris petiolaris* (two-sided mixed model associations; $n = 160$ individuals). The purple lines in **b**, **d** represent 5% Bonferroni-corrected significance. Only positions with −log10 *P* value >2 are plotted. **e**, Distribution of F_{ST} values for SNPs and haploblocks in comparisons between dune- and non-dune-adapted populations of *H. petiolaris fallax* in Texas and Colorado[16]. Percentiles are reported for the most highly divergent haploblocks. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. Number of individuals: $n = 28$ (Colorado); $n = 54$ (Texas). Number of SNPs: $n = 1{,}196{,}399$ (Colorado); $n = 1{,}169{,}273$ (Texas). **f**, Maximum-likelihood trees for two of the haploblocks segregating within *H. petiolaris*. Dune populations of *H. petiolaris fallax* are highlighted in light (Colorado) and dark tan (Texas). For pet09.01 and pet11.01, although both dune populations have converged on the same haplotype, the Texas haplotype is the ancestral *H. petiolaris fallax* copy, whereas in Colorado the haplotype is derived from introgression with *H. petiolaris petiolaris*, suggesting convergent adaptation. Bootstrap values for major nodes are reported (asterisks = 100).

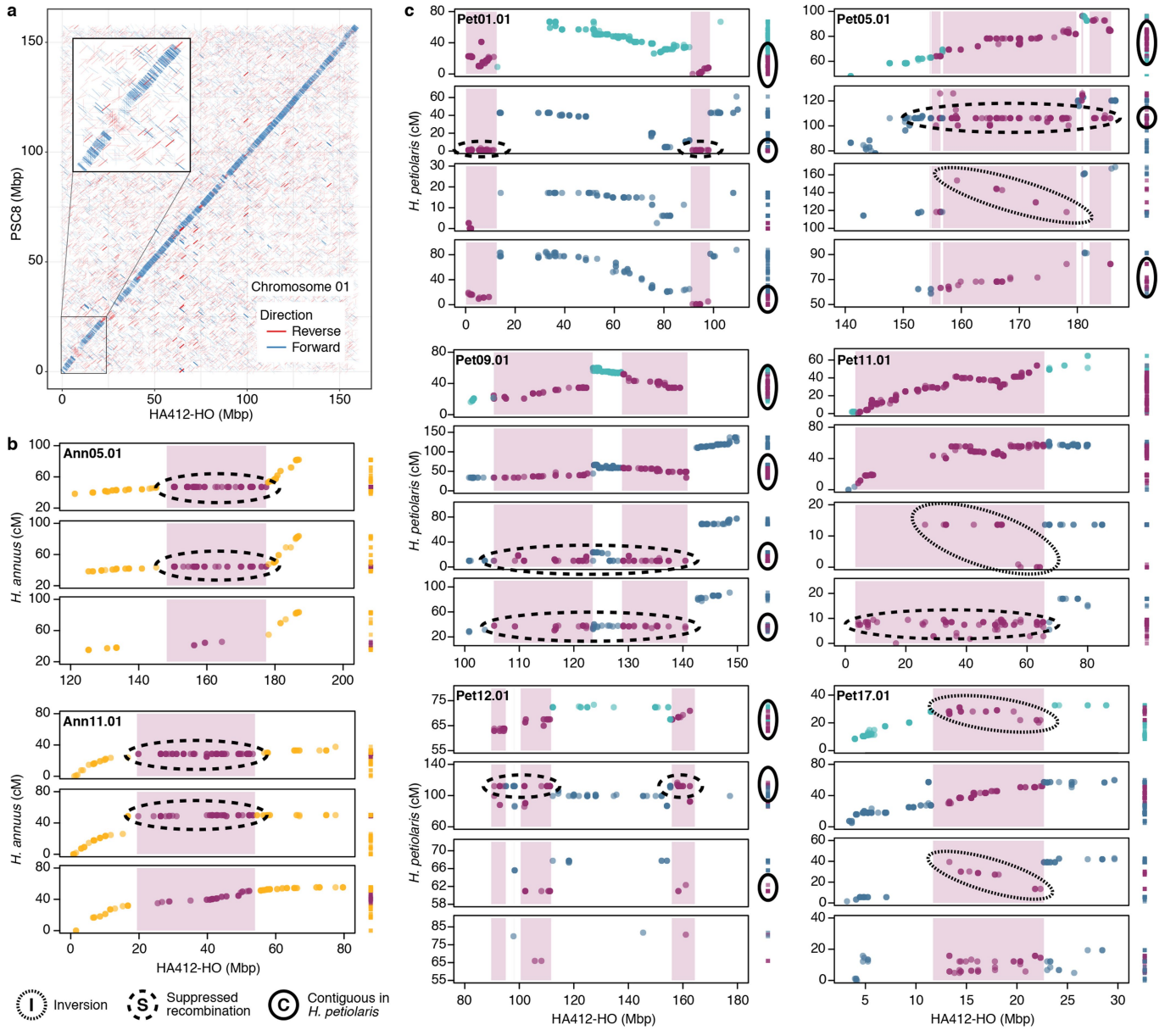


Extended Data Fig. 4 | See next page for caption.

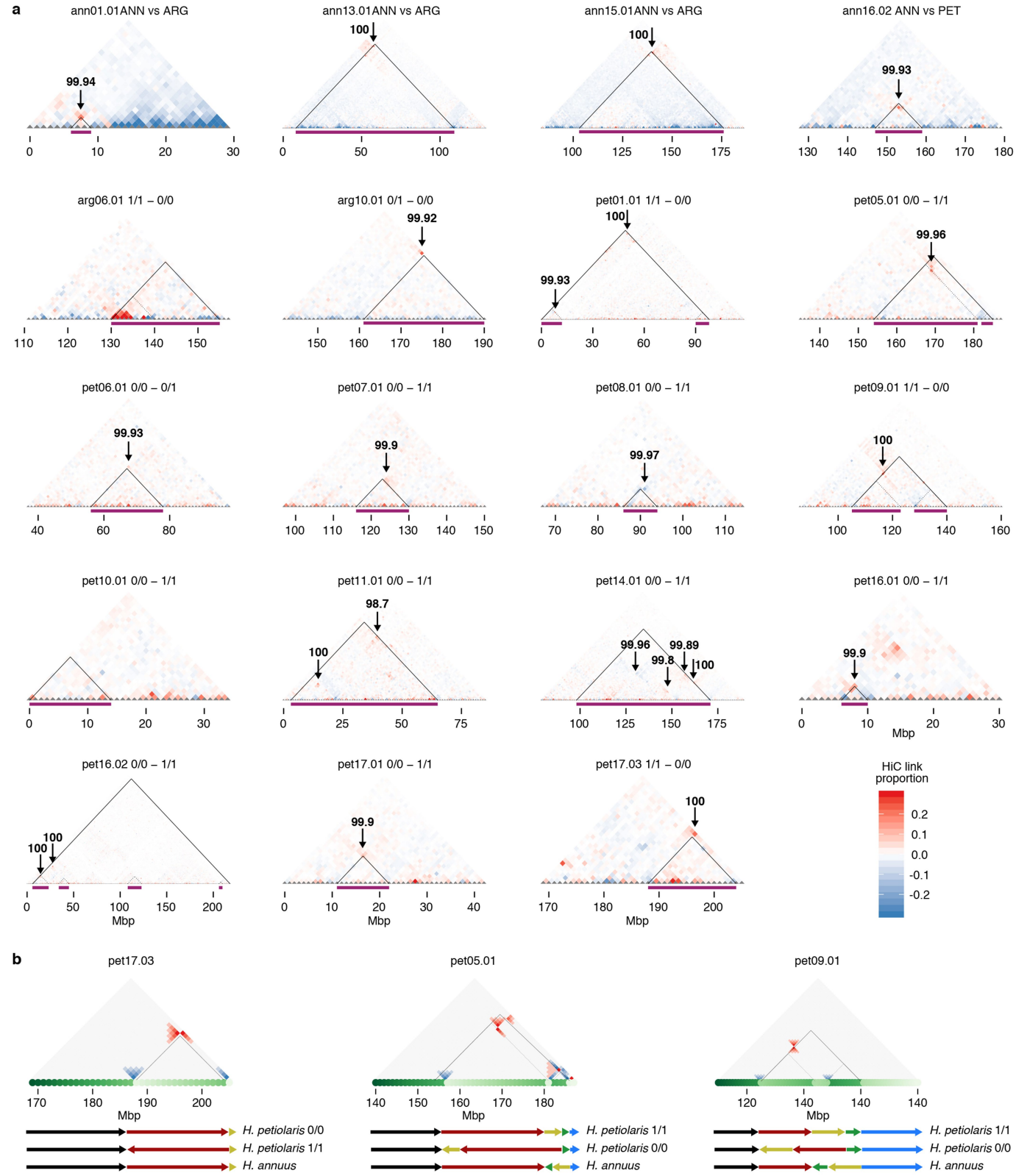
Extended Data Fig. 4 | Local PCA highlights haploblock regions. For each predicted haploblock, the local PCA MDS plot for the relevant chromosome, a PCA of the selected region, observed heterozygosity for each haploblock genotype and LD patterns for the relevant chromosome are shown. In the local PCA MDS plots, each dot represents a 100-SNP window, and windows within the haploblock region are highlighted. The *x*-axis values represent Mbp. For *H. petiolaris*, haploblocks were identified in the full species or subspecies datasets; the local PCA and LD plots are from the dataset in which the haploblock was identified, and PCA and heterozygosity plots use the full dataset. In PCA plots, samples are coloured by inferred haploblock genotype. For LD plots, upper triangle = all individuals; lower triangle = only individuals homozygous for the more common haploblock allele. Colours represent the second highest R^2 value in 0.5-Mbp windows. For most haploblock regions, high LD is driven by differences between haplotypes, so high LD is removed when only one haplotype is present. Box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges. Sample size for all haploblock analyses is provided in the Source Data, available at https://github.com/owensgl/haploblocks/.



Extended Data Fig. 5 | Geographical distribution of haploblock genotypes. Map showing collection locations of the three sunflower species and the frequency of haploblock genotypes at each collection site.

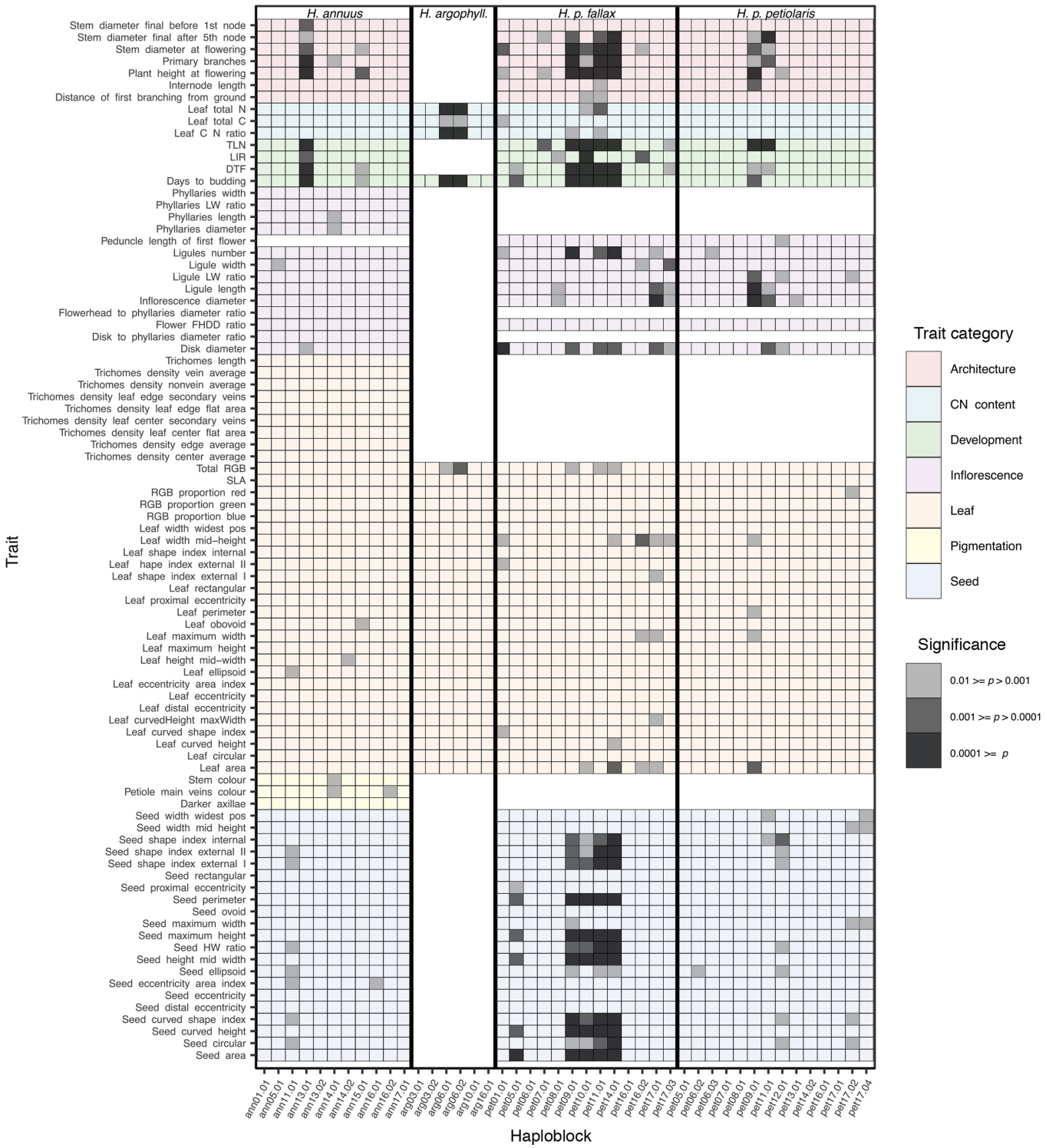


Extended Data Fig. 6 | Comparisons between reference assemblies and genetic maps confirm structural rearrangements associated with haploblocks. a, Alignment of chromosome 1 for the *H. annuus* genome assemblies PSC8 and HA412-HOv2. The ann01.01 region (at about 8 Mpb; inset), for which the two cultivars have different haplotypes, shows inverted alignment. **b**, Three *H. annuus* genetic maps (constructed using F_2 populations between wild individuals and the HA412-HO cultivar). **c**, Four genetic maps (constructed using F_1 populations. From top to bottom: *H. petiolaris petiolaris*, *H. petiolaris fallax*[87], newly constructed dune *H. petiolaris fallax* and newly constructed non-dune *H. petiolaris fallax*[88]) are plotted relative to the HA412-HOv2 reference assembly. To the right of each dot plot, markers are plotted in the order in which they appear in each genetic map. Haploblock regions and the markers that fall within them are highlighted in purple. Circled haploblock regions show evidence of different orientations across the multiple maps (dotted lines), of suppressed recombination (dashed lines) or are contiguous in *H. petiolaris* maps despite being split over multiple windows in the HA412-HOv2 reference assembly (solid lines). Parental haploblock genotypes are known for the *H. annuus* maps and for the bottom two *H. petiolaris* maps. Ann05.01 and ann11.01 were segregating within in the *H. annuus* mapping populations. Genotypes at pet05.01 and pet11.01 differed between the *H. petiolaris fallax* parents of newly constructed dune and non-dune populations, whereas both parents were heterozygous for the pet09.01 haploblock. In all these cases, patterns of segregation are consistent with the parental haploblock genotypes. For the remaining *H. petiolaris* maps, the parental haploblock genotypes are not known. Because an absence of evidence is uninformative in these cases, only haploblock regions with evidence for inversions or contiguous windows from these two maps are plotted.

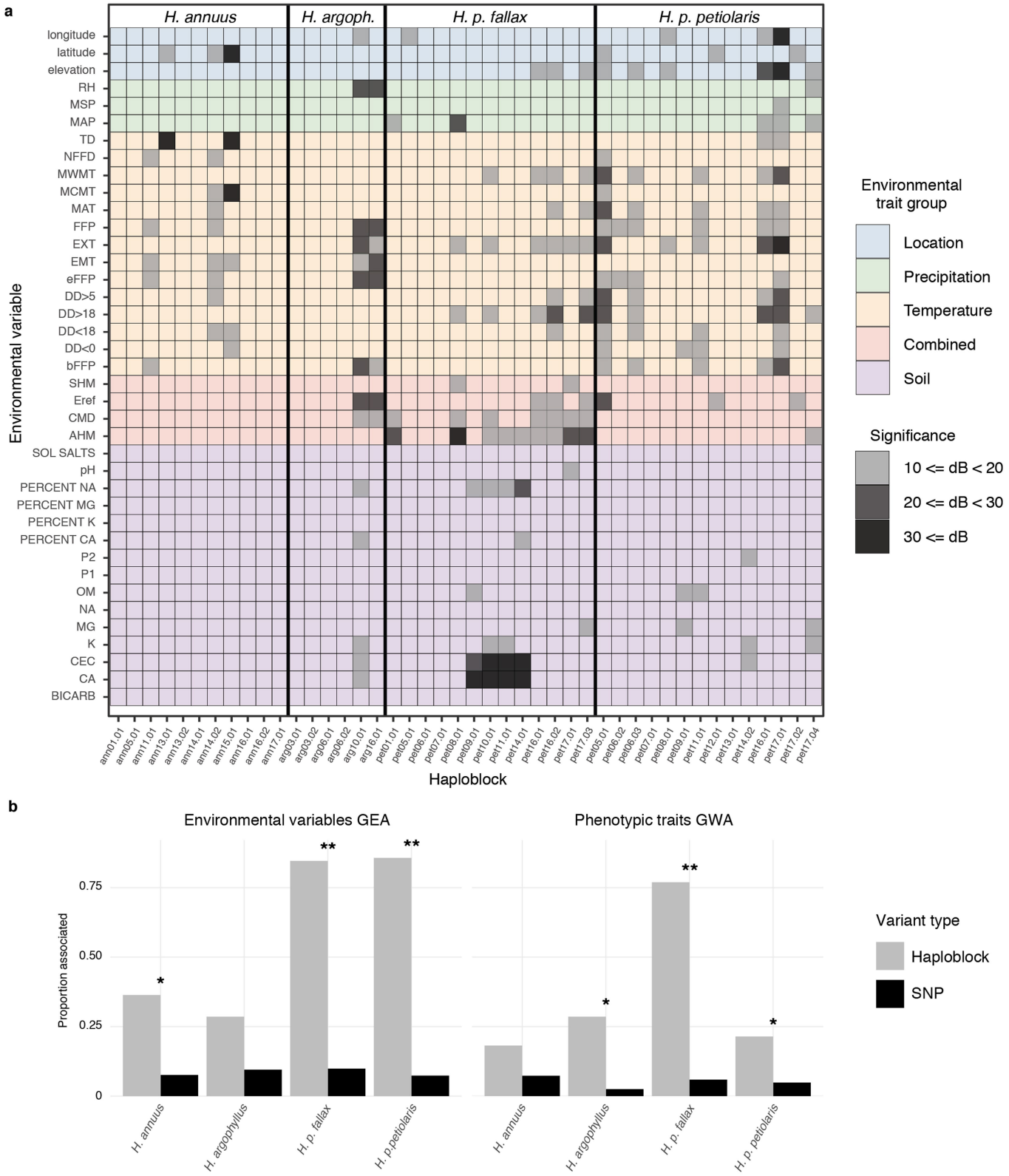


Extended Data Fig. 7 | See next page for caption.

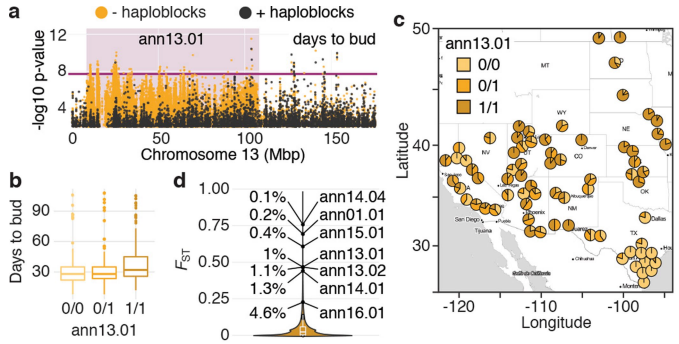
Extended Data Fig. 7 | HiC comparisons identify SVs associated with most, but not all, haploblocks. a, Differences in HiC interactions between pairs of early- and late-flowering *H. argophyllus* or dune and non-dune *H. petiolaris* samples. Purple bars and solid black lines represent approximate haploblock boundaries. Pieces of a single haploblock that map to different regions of the HA412-HOv2 reference are highlighted by dotted lines. Top row, comparisons between *H. annuus* and *H. argophyllus* or *H. petiolaris*, for *H. annuus* haploblock regions. Because the relative haploblock genotypes between sunflower species are not known, only cases in which evidence of structural variants were observed are reported. Following rows, regions for which the pairs of *H. argophyllus* or *H. petiolaris* samples differed at haploblock alleles. Red or blue dots show increased or decreased, respectively, long-distance interactions in one sample, consistent with differences in genome structure. Relevant differences in long-distance interactions are highlighted by black arrows; for each of these, the percentage rank compared to all other possible interactions at the same distance across the genome is reported. No evidence of large-scale structural variation was observed for arg06.01 and pet10.01. An excess of interactions in the early-flowering allele for the approximately 130–140-Mbp region of chromosome 6 is consistent with the presence of deletions in the late-flowering alleles (Extended Data Fig. 2c), as well as with improved mappability of reads from the early-flowering allele, which—being an introgression from wild *H. annuus*—is closer in sequence to the HA412-HO reference. Differences in HiC interactions were capped between −0.3 and 0.3 for plotting purposes. **b**, Inversion scenarios with comparisons of simulated HiC interaction matrixes consistent with empirical patterns. There are *H. annuus*-specific inversions in the reference genome, as well as inversions between haploblocks.



Extended Data Fig. 8 | Haploblock GWAs. Heat map of GWAs for individual phenotypic traits, treating haploblocks as individual loci. Haploblocks were filtered to retain only regions with minor allele frequency ≥ 3%. PCA and kinship matrices used as covariates were calculated without variants inside haploblock regions. GWAs were calculated using two-sided mixed models. Number of individuals: $n = 614$ (*H. annuus*); $n = 294$ (*H. argophyllus*); $n = 209$ (*H. petiolaris fallax*); $n = 163$ (*H. petiolaris petiolaris*).



Extended Data Fig. 9 | Haploblock GEAs. a, Heat map of GEAs for individual environmental variables, treating haploblocks as individual loci. Haploblocks were filtered to retain only regions with minor allele frequency ≥ 3%. The population correlation matrix was calculated without variants inside haploblock regions. GEAs were calculated using two-sided XtX statistics. Number of populations: $n = 71$ (*H. annuus*); $n = 30$ (*H. argophyllus*); $n = 23$ (*H. petiolaris fallax*); $n = 17$ (*H. petiolaris petiolaris*). **b**, The proportion of haploblock and SNP loci significantly associated with one or more environmental variable (dB ≥ 10) or phenotypic trait ($P \leq 0.001$). *$P < 0.05$, **$P < 0.0005$ (two-sided proportion test; exact P values and number of individuals are reported in Source Data).



Extended Data Fig. 10 | A 100-Mbp haploblock is associated with early flowering in the *texanus* ecotype of *H. annuus*. **a**, GWA for flowering in *H. annuus* (two-sided mixed model associations; $n = 612$ individuals), using a kinship matrix and PCA covariate including (black dots) or excluding (yellow dots) the haploblock regions. Haploblock regions are highlighted in purple. The purple line represents 5% Bonferroni-corrected significance. Only positions with $-\log_{10} P$ value >2 are plotted. **b**, Flowering time for individuals with different genotypes at ann13.01. Number of individuals: $n = 244$ (0/0); $n = 168$ (0/1); $n = 200$ (1/1). **c**, Distribution of ann13.01 haplotypes. **d**, Distribution of F_{ST} values for individual SNPs and haploblocks in comparisons between the *texanus* ecotype of *H. annuus* and other *H. annuus* populations. Percentiles are reported for the most highly divergent haploblocks. In **b**, **d**, box plots show the median, box edges represent the 25th and 75th percentiles, whiskers represent the maximum and minimum data points within 1.5× interquartile range outside box edges.

Extended Data Table 1 | Positions and frequencies of haploblocks, and experimental support for linked SVs.

Species	Haploblock ID	Chr.	Allele freq.	Hi-C genotype	Support	Size (Mbp)	Region (Mbp)
H. annuus	ann01.01	1	0.21	0/0	A, h	4	6-10
H. annuus	ann05.01	5	0.52	0/0	A, g	29	148-177
H. annuus	ann11.01	11	0.25	0/0	g	35	19-54
H. annuus	ann13.01	13	0.62	1/1	h, g	101	9-110
H. annuus	ann13.02	13	0.13	0/0	g	19	139-158
H. annuus	ann14.01	14	0.22	0/0		28	101-129
H. annuus	ann14.02	14	0.9	1/1		6	129-135
H. annuus	ann15.01	15	0.31	0/0	h, g	72	104-176
H. annuus	ann16.01	16	0.85	1/1		13	10-23
H. annuus	ann16.02	16	0.96	1/1	h, g	12	147-159
H. annuus	ann17.01	17	0.81	1/1		9	188-197
H. argophyllus	arg03.01	3	0.89	1/1 : 1/1		47	42-89
H. argophyllus	arg03.02	3	0.06	0/0 : 0/0		14	28-42
H. argophyllus	arg06.01	6	0.08	1/1 : 0/0	No SV	25	130-155
H. argophyllus	arg06.02	6	0.05	0/0 : 0/0		5	125-130
H. argophyllus	arg10.01	10	0.19	0/1 : 0/0	H	29	161-190
H. argophyllus	arg16.01	16	0.14	0/0 : 0/0		3	202-205
H. petiolaris	pet01.01	1	0.67; 0.97	1/1 : 0/0	H, g	20	0-12, 91-99
H. petiolaris	pet05.01	5	0.91; 0.62	0/0 : 1/1	H, G, g	28	154-186
H. petiolaris	pet06.01	6	0.23; 0	0/0 : 0/1	H	23	56-79
H. petiolaris	pet06.02	6	0; 0.11	0/0 : 0/0	h	37	19-56
H. petiolaris	pet06.03	6	0; 0,16	0/0 : 0/0		2	9-10
H. petiolaris	pet07.01	7	0.59; 0.08	0/0 : 1/1		14	116-130
H. petiolaris	pet08.01	8	0.70; 0.07	0/0 : 1/1	H	7	87-94
H. petiolaris	pet09.01	9	0.17; 0.36	1/1 : 0/0	H, g	32	105-123, 128-141
H. petiolaris	pet10.01	10	0.86; 1	0/0 : 1/1	No SV	14	0-14
H. petiolaris	pet11.01	11	0.78; 0.39	0/0 : 1/1	H, G, g	62	3-65
H. petiolaris	pet12.01	12	0; 0.09	0/0 : 0/0	g	26	89-95, 100-111, 155-164
H. petiolaris	pet13.01	13	0.99; 0.65	1/1 : 1/1		1	148-149
H. petiolaris	pet14.01	14	0.88; 0.99	0/0 : 1/1	H	73	98-171
H. petiolaris	pet14.02	14	0.01; 0.68	0/0 : 0/0		9	135-144
H. petiolaris	pet16.01	16	0.74; 0.36	0/0 : 1/1	H	4	6-10
H. petiolaris	pet16.02	16	0.79; 1	0/0 : 1/1	H	53	6-23,34-46, 109-112, 118-123, 137-149, 206-210, 214-218
H. petiolaris	pet17.01	17	0.62; 0.16	0/0 : 1/1	H, G	11	12-23
H. petiolaris	pet17.02	17	1; 0.91	1/1 : 1/1		29	39-68
H. petiolaris	pet17.03	17	0.29; 0	1/1 : 0/0	H	17	188-205
H. petiolaris	pet17.04	17	1; 0.90	1/1 : 1/1		11	194-205

Allele frequencies for haplotype 1 are reported. For *H. petiolaris,* allele frequencies for *H. p. fallax* and *H. p. petiolaris,* respectively, are reported. Hi-C genotype, haploblock genotypes of the pair of individuals that were used for Hi-C sequencing (for *H. annuus* haploblocks, only the genotype of HA412-HO is reported). Experimental support for haploblock: H, differences in Hi-C patterns between samples; h, differences in Hi-C patterns relative to the HA412-HO reference; A, differences between reference *H. annuus* assemblies; G, differences in orientation between genetic maps; g, recombination suppression in genetic maps; no SV, no evidence of SV in Hi-C experiments despite appropriate comparison. Haploblock positions are relative to the HA412-HOv2 assembly; in some cases predicted haploblocks are in multiple pieces owing to rearrangements relative to the reference.

nature research

Corresponding author(s): Loren H. Rieseberg
Natalia Bercovich

Last updated by author(s): Mar 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ImageJ (v2.0.0) and Tomato Analyzer (v3.0) were used to extract morphometric data from digital images
Data analysis	The following software were used to process short read data and for variant calling: Trimmomatic (v0.36), NextGenMap (v0.5.3), samtools (v0.1.19), picard (v2.9.3), sambamamba (v0.6.6), GATK (v4.0.1.2), Snakemake (v4.7.0), BWA (v0.7.17), vcftools (v0.1.13). GWA analyses were performed using EMMAX (v07Mar2010) and easyGWAS (v2.9); Beagle (v10Jun18.811) was used for genotype imputation. GEA analyses were performed using BayPass (v2.1). Haploblock detection and analysis were performed using lostruct (v0.0.0.9) and SNPrelate. (v1.16.0) HiC data were analyzed using HOMER (v4.10). Genome assemblies comparisons were performed using MUMmer (v4.0.0b2) and Syri (v1.0). Phylogenetic analyses and haploblock dating were performed using IQtree (v1.6.10), R (v3.5.1), BEAST (v1.10.4), Tracer (v1.7.1) and Figtree (v1.4.4). R (v3.6.2) and R studio (v1.1.456) were used for most statistical analyses and for plotting data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequenced data are stored in the Sequence Read Archive (SRA) under BioProjects PRJNA532579, PRJNA398560 and PRJNA564337. Accession numbers for individual samples are listed in Supplementary Table 1 (tabs "Coverage and analyses", "Outgroups", "Samples from other studies" and "HiC samples"). The HA412-HOv2 and PSC8 genome assemblies are available at https://sunflowergenome.org/ and https://heliagene.org/. Filtered SNP datasets are available at https://rieseberglab.github.io/ubc-sunflower-genome/ and will be made public on April 12th, 2020. GWA results, as well as the corresponding SNP and trait data are available at https://easygwas.ethz.ch/gwas/myhistory/public/20/, https://easygwas.ethz.ch/gwas/myhistory/public/21/, https://easygwas.ethz.ch/gwas/myhistory/public/22/, https://easygwas.ethz.ch/gwas/myhistory/public/23/. HaFT1, HaFT2 and HaFT6 sequences have been deposited in GenBank under accession numbers MN517758-MN517761. Source data for figures 1a,b; 2b,e,g,h,l; 3c,d; 4b,c,d,g; 5a,b,e,f; and for Extended Data Figures 1d,e; 2a,f; 3a; 6b,c: 9b are provided with the paper. Source data for all figures and all code associated with this project are provided at https://github.com/owensgl/haploblocks/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Plants representing 151 populations of three species of wild sunflowers were grown in a common garden experiment. The genomes of 1506 individuals were re-sequenced, and used for genome-wide association, genotype-environment association analyses, phylogenetic analyses, and local PCA studies.
Research sample	Seeds and soil samples were collected from populations of wild sunflowers (Helianthus annuus, H. argophyllus and H. petiolaris), covering their entire native range in the USA and Canada. Ten plants from 151 of these populations (H. annuus = 71, H. argophyllus = 30, H. petiolaris = 50) were grown and phenotyped in a common garden experiment. The genomes of 1401 of these plants and 105 individuals from a previous experiments were re-sequenced.
Sampling strategy	Seeds from wild populations were collected from 21-37 randomly chosen individuals, from previously described or newly identified populations. Populations were selected to cover the natural range of those sunflower species. Ten individuals were grown and re-sequenced for each population because this would provide a good representation of the variation present in each population, while maximizing the number of populations that could be surveyed. The total number of samples per species (719 H. annuus, 488 H. petiolaris, 299 H. argophyllus) were sufficient to provide an 85% probability of detecting loci explaining 5% or more of the phenotypic variance in H. annuus, 8% of variance in H. petiolaris, and 12% of variance in H. argophyllus. For flowering time analyses in transgenic Arabidopsis thaliana lines, samples size was determined by the number of primary transformants that could be recovered (10-47).
Data collection	Seeds and soil samples from wild sunflower populations were collected by D.O.B. Three to five soil samples (0 - 25 cm depth) were collected with a corer at each population, from across the area in which seeds were collected. N.B., M.T. and I.I. collected phenotype data for the common garden individuals, with help from other co-authors. N.B and M.T. collected leaf samples, extracted DNA and generate whole-genome shotgun sequencing libraries. M.T. performed HaFT1 expression analyses and transgenic experiments.
Timing and spatial scale	Seed from wild populations were collected between Aug. 30th and Dec. 4th, 2015. Plants in the common garden experiment were planted starting on May, 25th 2016, and grown until the beginning of November 2016.
Data exclusions	Individuals were excluded from analyses when their phylogenetic placement, based on re-sequencing data, was not consistent with their presumed population or species of origin (signifying likely contamination or mis-labeling). This is a standard quality control strategy for population or evolutionary genomic studies in our lab and others for detecting mis-labeled or contaminated samples.
Reproducibility	Due to its size and complexity, the common garden and re-sequencing experiments were not replicated. At least three technical replicates and at least two biological replicates were done for HaFT1 expression analyses; all gave consistent results. At least 17 independent primary transformants were analyzed for each Arabidopsis transgenic line, and all displayed consistent phenotypes.
Randomization	In the common garden experiment, each sunflower species was grown in a separate field. Pairs of plants from the same population were randomly distributed within each field (plants were paired to facilitate within-population crosses).
Blinding	Researcher were not blinded as to the identity of individual samples. However, information about their populations of origin were not attached to the samples during data acquisition.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	Field work was limited to collection of seeds and soil samples from populations of wild sunflowers from throughout the USA and Canada. Twenty-four climatic factors and 15 soil properties for each population collection location are provided in Supplementary Table 1 (tab "Populations, environ. variables").
Location	Samples were collected throughout the USA and Canada. Location information for all populations, as well as description of the collection sites, are reported in Supplementary Table 1 (tab "Populations, environ. variables").
Access and import/export	Samples were largely collected on public land or with the permission of the land owner. Permits were obtained for collecting samples at the Welder Wildlife Refuge in Texas, USA (H. argophyllus), and at the Bear River Migratory Bird Refuge in Utah, USA (H. annuus). An import permit for seeds was secured from the Canadian Food Inspection Agency (CFIA)
Disturbance	Sampling was non-intrusive and did not produce any habitat disturbance.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging