Genomic islands of divergence are not affected by geography of speciation in sunflowers.

Renaut S[1,5], Grassa CJ[1], Yeaman S[1], Moyers BT[1], Lai Z[2], Kane NC[1,3], Bowers JE[4], Burke JM[4], Rieseberg LH[1,2].

Supplementary Online Material

1. Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

2. Department of Biology, Center for Genomics and Bioinformatics, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA

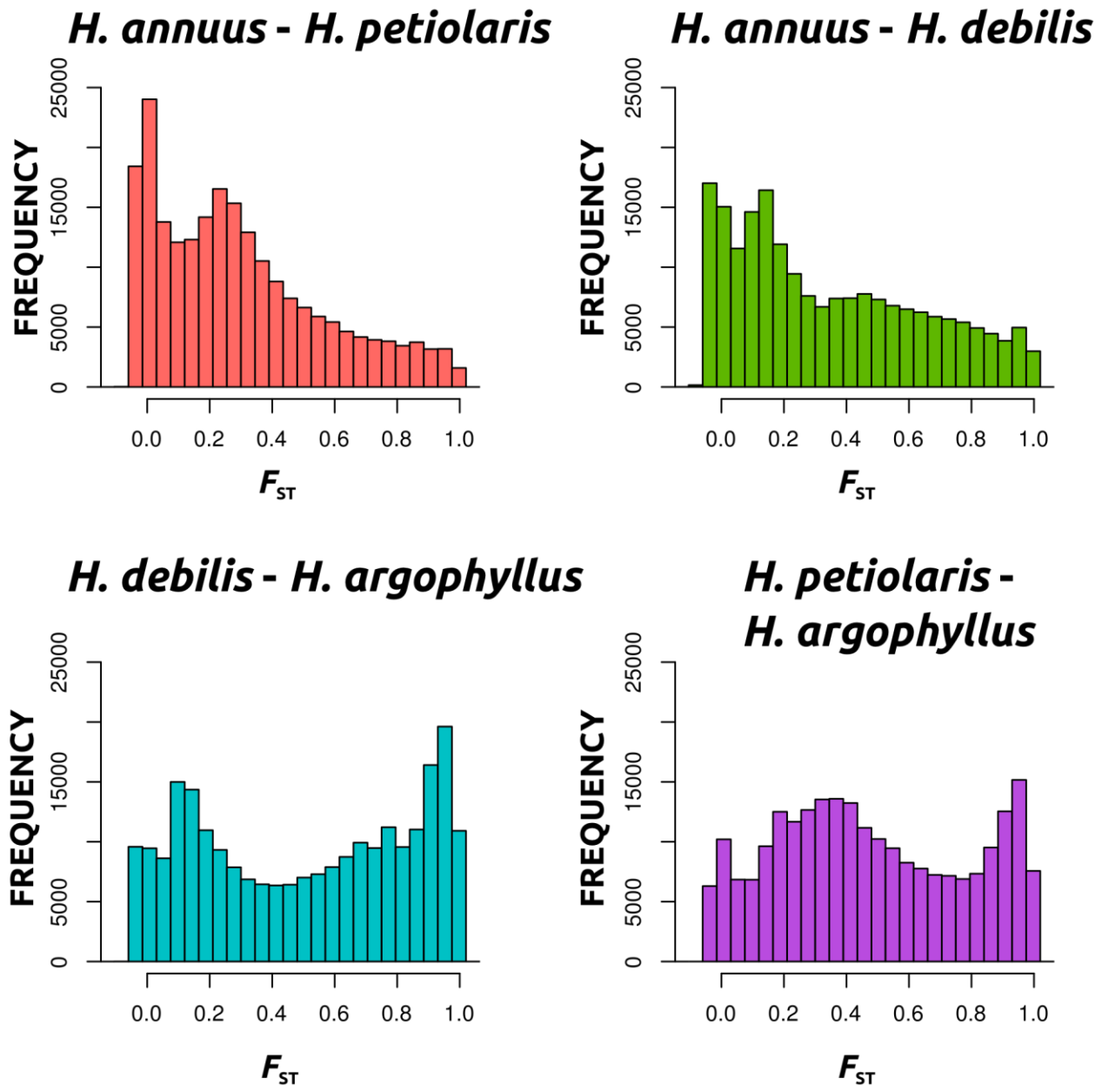
3. Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309, USA

4. Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

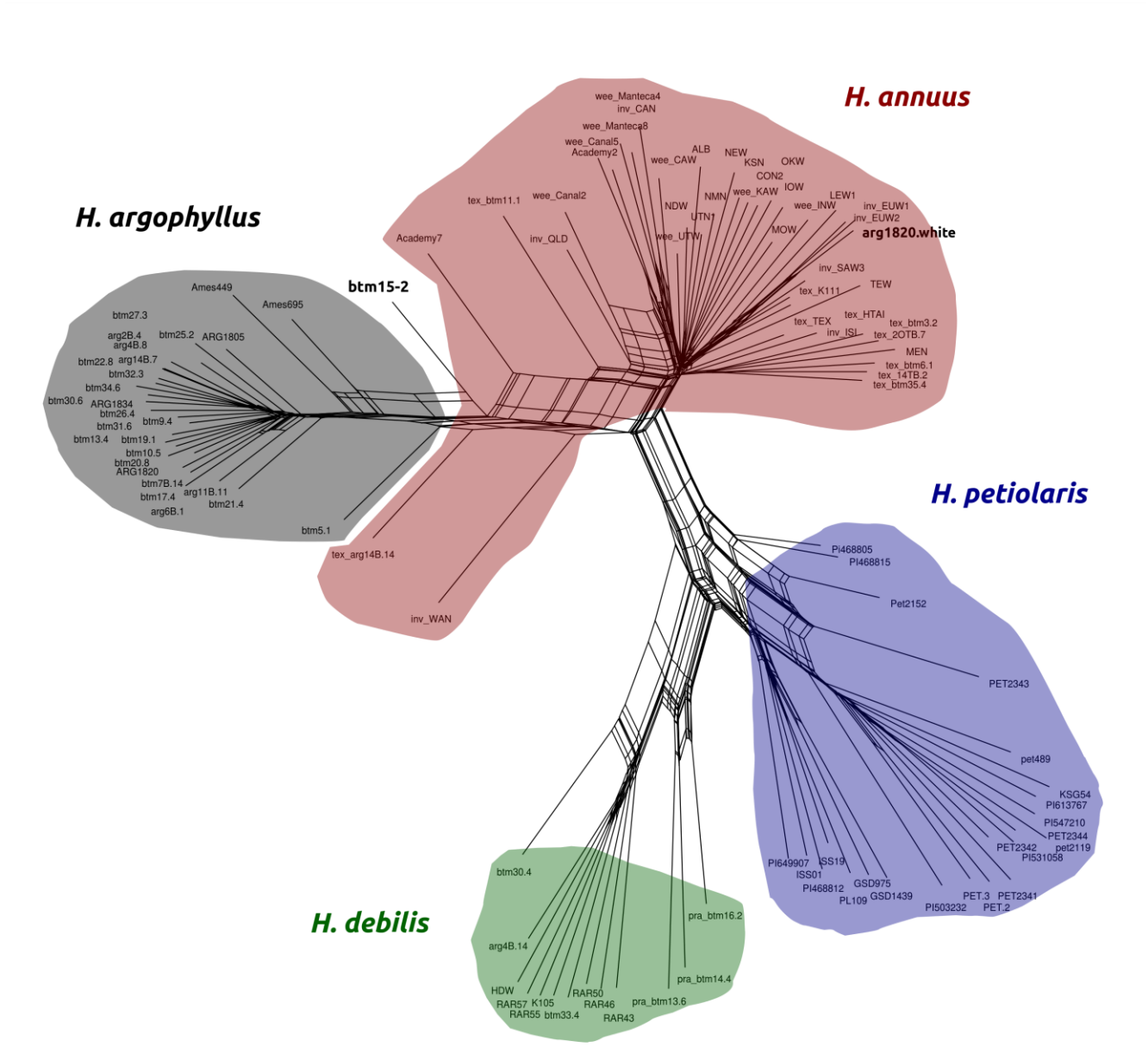
5. Corresponding author: sebastien.renaut@gmail.com; Phone (604) 827-3535; Fax (604) 822-6089

Supplementary Figures

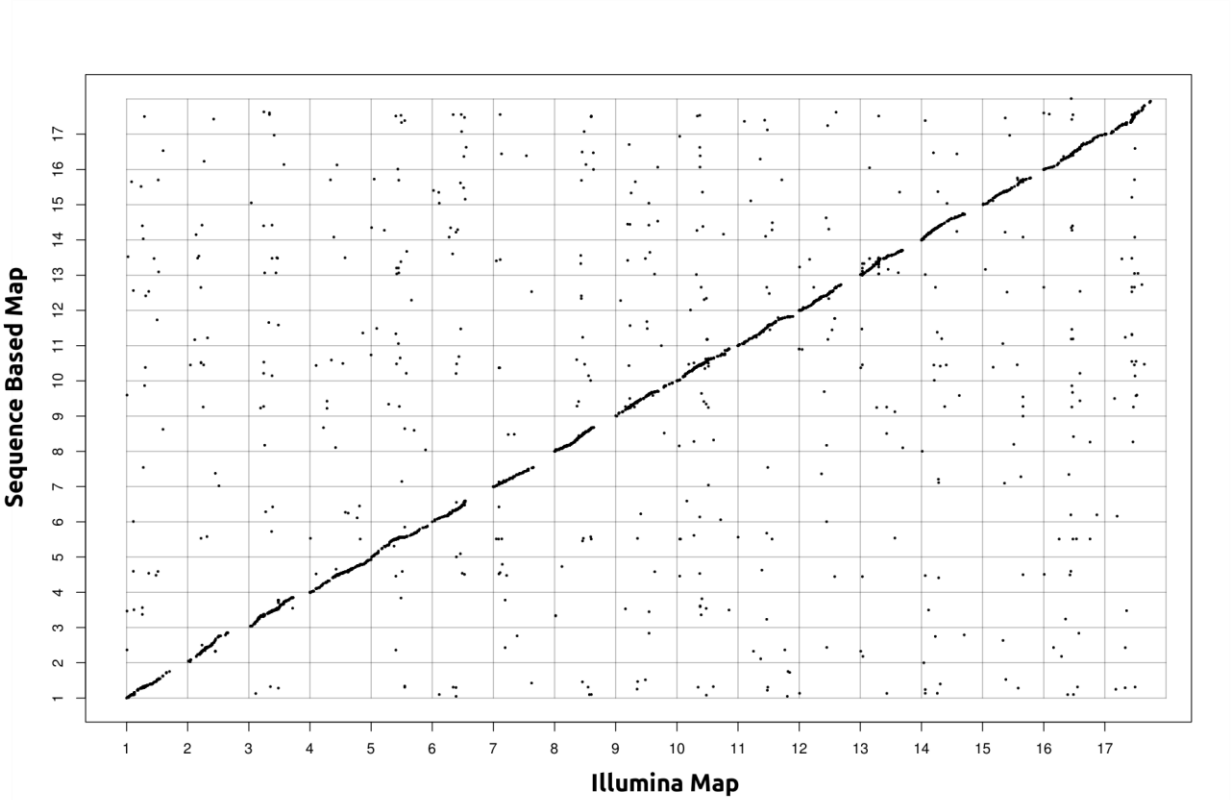
Supplementary Figure S1 | F_{ST} distribution for each species pair. The four panels show the F_{ST} distribution for the four considered species pair.



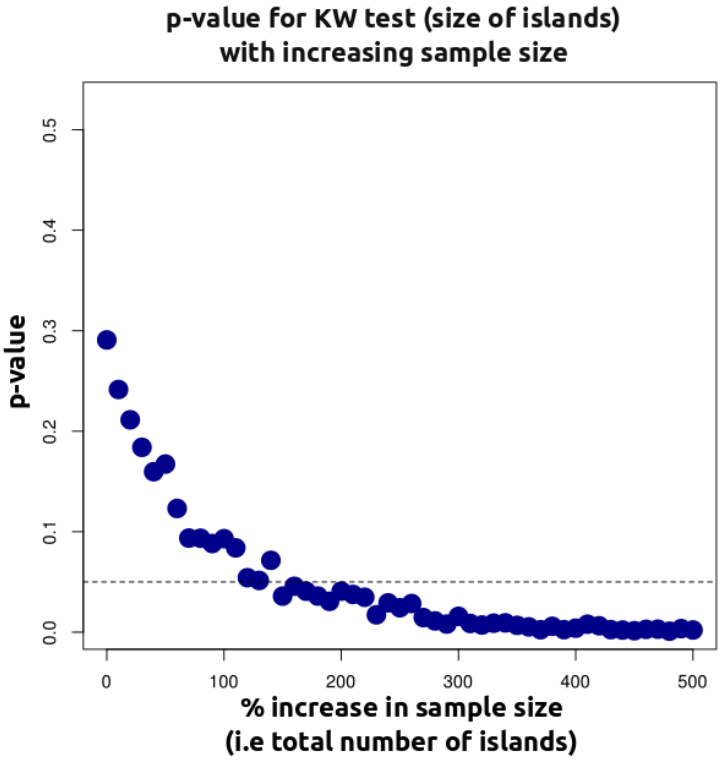
Supplementary Figure S2 | Phylogenetic network reconstructed with the Neighbour-net method. The network is based on 10,000 high quality SNPs for 105 individuals (two individuals were not included in the network, see *Bioinformatics* section in methods for rationale). Individuals *ARG1820.white* and *btm15-2* were also removed from further analysis given their unexpected phylogenetic relationship.



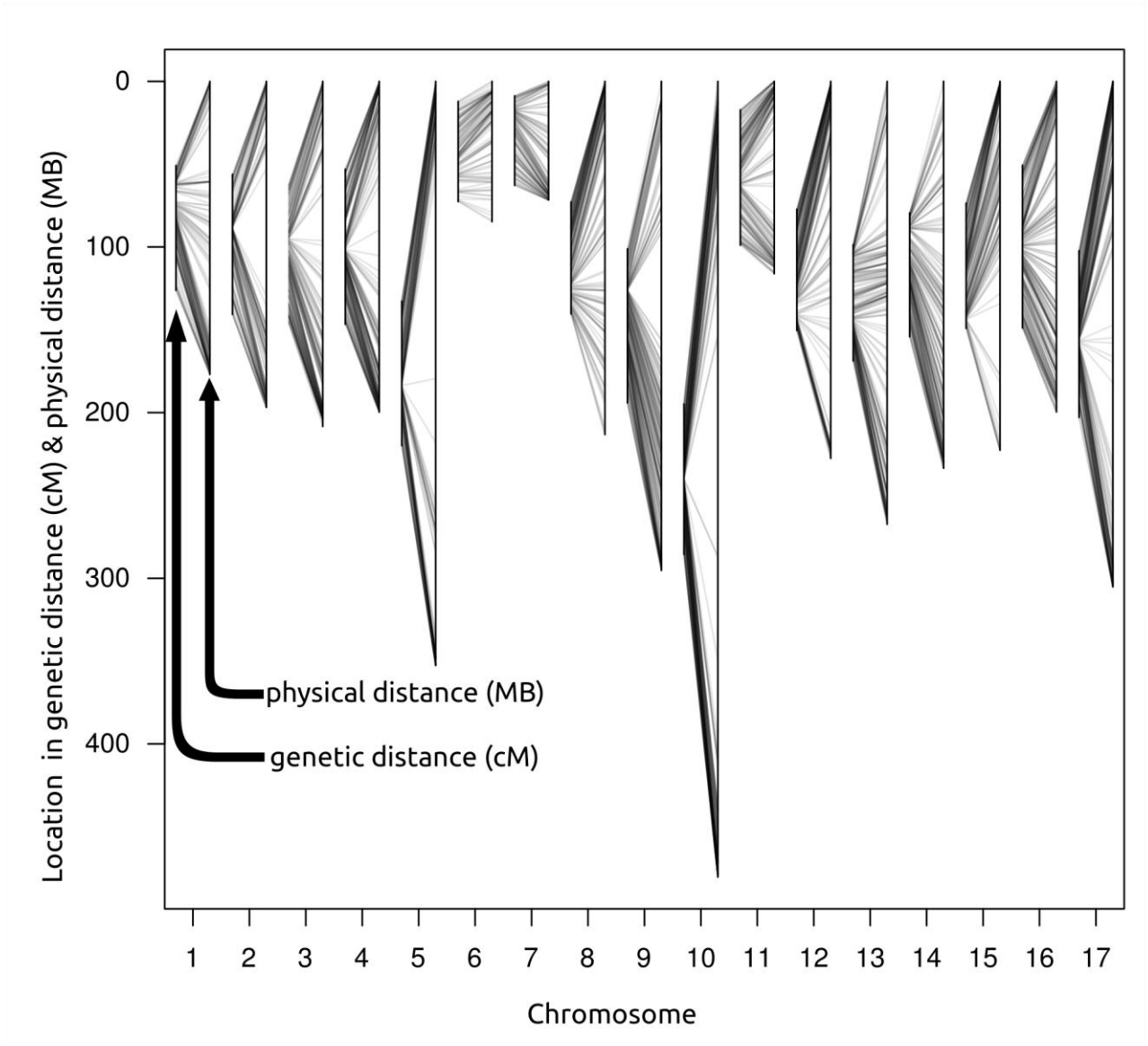
Supplementary Figure S3 | Synteny between genetic maps. Comparison of synteny between the 4,727 markers found in both the map generated for this study (sequence based map) and a completely independently constructed 10,800 locus (Illumina) map of Bowers and colleagues[53]. Off diagonal points likely represent SNPs in paralogous loci that map to different locations in different mapping populations.



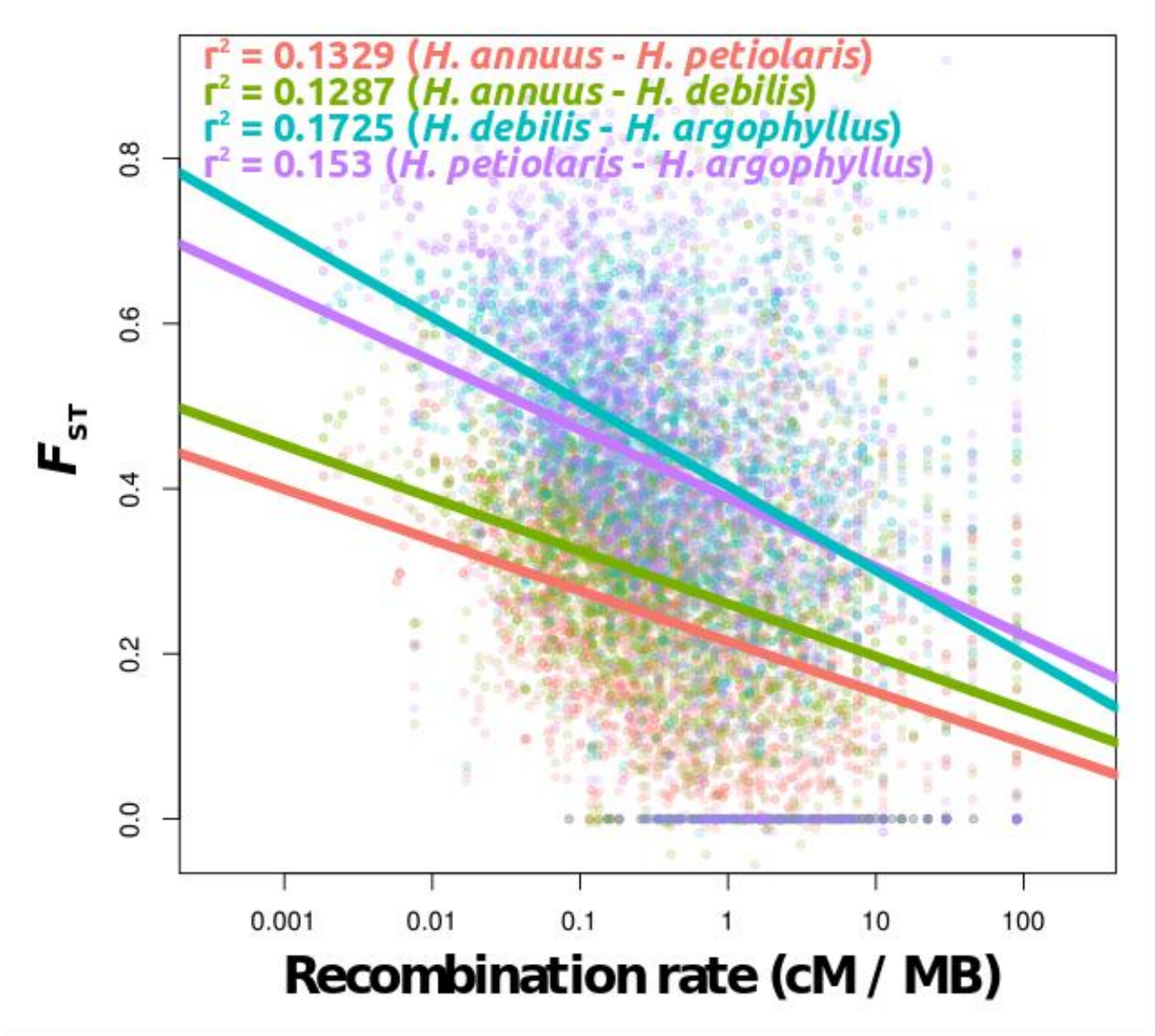
Supplementary Figure S4 | Effects of sample size on detection of differences in island numbers. Sample sizes required to detect significant differences in island size among the four species pairs. We determined the sample sizes needed to detect significant differences in island sizes by artificially increasing the number of islands until the test statistic became significant. We did this by augmenting in 5% increments (through a resampling approach, with replacement and within a species pair) the number of islands identified in each comparison until Kruskal-Wallis tests became significant. Sample sizes would have to be nearly twice as large for differences in island sizes to be significantly different (dashed line, *p*-value < 0.05).



Supplementary Figure S5 | Comparison of genetic map and physical map. Genetic map; 24,406 contigs positioned unto 3,047 unique map positions, mean distance between map positions = 0.45, mean number of SNP per map positions = 62, mean number of contigs per map position = 9, total size = 1,371 centimorgans. Physical distances (megabases) are based on the placement of physical map contigs onto the genetic map. Large restrictions in recombination distance (e.g., chromosomes 5 and 10) likely represent centromeric regions, which are known to vary in size between chromosomes.

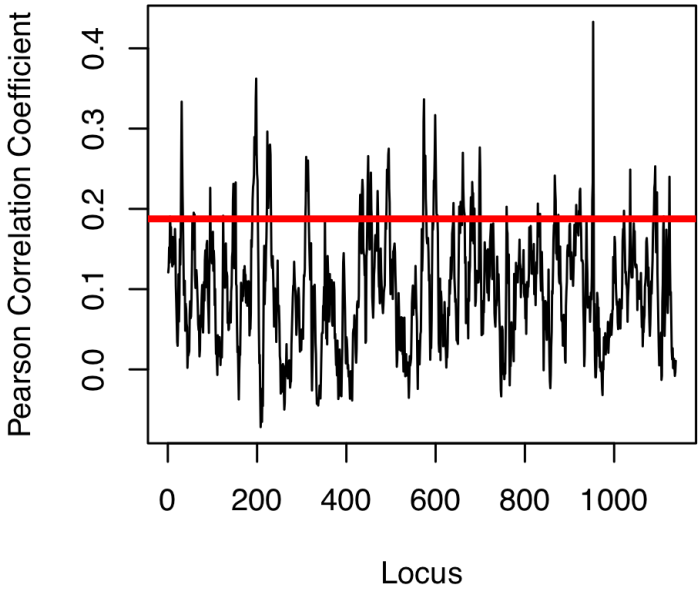


Supplementary Figure S6 | Effect of recombination rate on genetic divergence. Correlations between recombination rates and mean F_{ST} for all four comparisons.

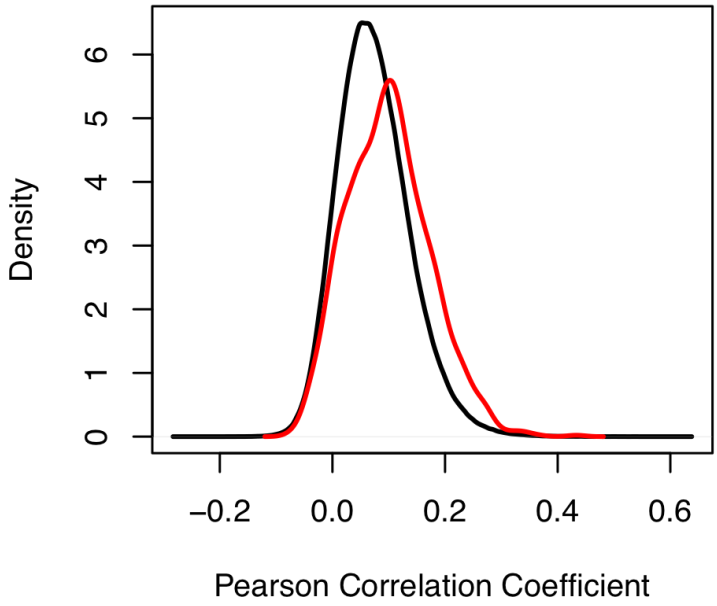


Supplementary Figure S7 | Co-expression among neighbouring genes. (A) Pearson correlation coefficients calculated within an 11-gene sliding window (chromosome 1). The *x*-axis shows the ordered position of the loci along the chromosome. Horizontal red line in top panel shows the 95th percentile of the permuted distribution. (B) Probability density for the permuted distribution (black) and the observed distribution (red) of Pearson correlation coefficients for co-expression.

(A)



(B)



Supplementary Tables

Supplementary Table S1 | Summary of sequencing statistics and GPS locations of samples. Total number of sequences = 1.487 G, Total number of nucleotide = 298.7 GB **(**M = 10e6, G = 10e9**,** B = base)

sample name	species	sequencing technology	number of sequences (M)	normali-zation	location	latitude	longitude
Academy2	annuus	illumina	8.74	Non-normalized	California	-119.56	36.74
Academy7	annuus	illumina	13.51	Non-normalized	California	-119.56	36.74
ALB	annuus	illumina	11.28	normalized	Alberta	-115.00	54.67
LEW1	annuus	illumina	12.73	normalized	Mexico	-109.82	27.38
MEN	annuus	illumina	8.62	normalized	USA	-	-
NEW	annuus	illumina	11.57	normalized	Nebraska	-97.67	41.37
TEW	annuus	illumina	10.44	normalized	Tennessee	-83.92	35.96
Ames449	argophyllus	illumina	17.93	Non-normalized	Australia	150.83	-23.24
Ames695	argophyllus	illumina	17.02	Non-normalized	North Carolina	-78.00	33.88
arg11B-11	argophyllus	illumina	14.48	Non-normalized	Texas	-97.80	27.27
arg14B-7	argophyllus	illumina	16.75	Non-normalized	Texas	-97.39	28.70
ARG1805	argophyllus	illumina	16.13	Non-normalized	Texas	-97.40	27.78
ARG1820	argophyllus	illumina	22.04	Non-normalized	Texas	-98.13	26.89
ARG1834	argophyllus	illumina	16.36	Non-normalized	Texas	-97.01	28.81
arg2B-4	argophyllus	illumina	12.26	Non-normalized	Texas	-97.75	29.36
arg4B-8	argophyllus	illumina	20.68	Non-normalized	Texas	-97.14	28.73
arg6B-1	argophyllus	illumina	19.27	Non-normalized	Texas	-97.03	28.11
btm10-5	argophyllus	illumina	10.62	Non-	Texas	-97.14	27.92

				normalized			
btm13-4	argophyllus	illumina	19.92	Non-normalized	Texas	-97.04	28.04
btm17-4	argophyllus	illumina	14.74	Non-normalized	Texas	-97.29	27.45
btm19-1	argophyllus	illumina	17.05	Non-normalized	Texas	-97.22	27.62
btm20-8	argophyllus	illumina	14.3	Non-normalized	Texas	-97.17	27.68
btm21-4	argophyllus	illumina	19.11	Non-normalized	Texas	-97.12	27.76
btm22-8	argophyllus	illumina	15.93	Non-normalized	Texas	-97.08	27.86
btm25-2	argophyllus	illumina	17.27	Non-normalized	Texas	-97.31	28.27
btm26-4	argophyllus	illumina	11.83	Non-normalized	Texas	-97.33	28.44
btm27-3	argophyllus	illumina	14.19	Non-normalized	Texas	-97.33	28.62
btm30-6	argophyllus	illumina	13.12	Non-normalized	Texas	-97.50	29.20
btm31-6	argophyllus	illumina	19.93	Non-normalized	Texas	-97.66	29.27
btm32-3	argophyllus	illumina	13.83	Non-normalized	Texas	-97.69	29.65
btm34-6	argophyllus	illumina	19.77	Non-normalized	Texas	-97.78	29.55
btm5-1	argophyllus	illumina	14.59	Non-normalized	Texas	-98.15	27.14
btm7B-14	argophyllus	illumina	17.78	Non-normalized	Texas	-97.89	27.26
btm9-4	argophyllus	illumina	21.85	Non-normalized	Texas	-97.19	27.87
arg4B-14	debilis	illumina	27.68	Non-normalized	Texas	-97.14	28.73
btm30-4	debilis	illumina	26.23	Non-normalized	Texas	-97.50	29.20
btm33-4	debilis	illumina	18.17	Non-normalized	Texas	-97.70	29.53
K105	debilis	illumina	30.52	normalized	Texas	98.00	30.00
RAR43	debilis	illumina	26.41	normalized	Texas	-99.07	28.85
RAR46	debilis	illumina	24.64	normalized	Texas	-98.28	29.06
RAR50	debilis	illumina	20.11	normalized	Texas	-98.10	29.43
RAR55	debilis	illumina	18.14	normalized	Texas	-97.38	30.10
RAR57	debilis	illumina	20.86	normalized	Texas	-97.36	30.32
GSD1439	petiolaris	illumina	15.04	normalized	Colorado dunes	-105.64	37.73
GSD975	petiolaris	illumina	12.7	normalized	Colorado	-105.64	37.73

ISS01	petiolaris	illumina	2.79	normalized	Texas	-102.89	31.59
ISS19	petiolaris	illumina	10.87	normalized	Texas	-102.89	31.59
KSG54	petiolaris	illumina	9.81	normalized	Kansas	-98.77	38.37
pet2119	petiolaris	illumina	27.64	Non-normalized	Montana	-105.70	46.38
Pet2152	petiolaris	illumina	10.98	Non-normalized	Colorado	-104.82	40.82
PET-2	petiolaris	illumina	16.83	Non-normalized	Colorado	-105.60	39.06
PET2341	petiolaris	illumina	23.33	Non-normalized	Manitoba	-100.68	49.39
PET2342	petiolaris	illumina	19.37	Non-normalized	North Dakota	-97.45	48.72
PET2343	petiolaris	illumina	23.1	Non-normalized	Colorado	-102.25	40.14
PET2344	petiolaris	illumina	24.77	Non-normalized	North Dakota	-96.90	46.15
PET-3	petiolaris	illumina	15.52	Non-normalized	Colorado	-105.60	39.06
pet489	petiolaris	illumina	24.25	Non-normalized	Texas	-102.33	35.68
Pi468805	petiolaris	illumina	6.35	Non-normalized	New Mexico	-108.27	32.77
PI468812	petiolaris	illumina	18.51	Non-normalized	New Mexico	-103.34	33.51
PI468815	petiolaris	illumina	6.91	Non-normalized	Utah	-112.53	37.05
PI503232	petiolaris	illumina	26.18	Non-normalized	New Jersey	-75.04	39.41
PI531058	petiolaris	illumina	15.2	Non-normalized	North Dakota	-96.97	46.53
PI547210	petiolaris	illumina	19.09	Non-normalized	Illinois	-90.43	40.02
PI586932b	petiolaris	illumina	6.88	Non-normalized	Nebraska	-100.38	42.10
PI613767	petiolaris	illumina	27.92	Non-normalized	Oklahoma	-97.81	36.67
PI649907	petiolaris	illumina	26.61	Non-normalized	Texas	-103.19	32.12
PL109	petiolaris	illumina	1.28	normalized	New Mexico	-104.63	34.86
btm13-6	debilis	illumina	24.32	Non-normalized	Texas	-97.04	28.04
btm14-4	debilis	illumina	19.68	Non-normalized	Texas	-97.04	28.09
btm15-2	debilis	illumina	33.69	Non-normalized	Texas	-97.28	27.49
btm16-2	debilis	illumina	23.92	Non-	Texas	-97.29	27.43

				normalized			
Canal2	annuus	illumina	15.84	Non-normalized	California	-119.41	36.72
Canal5	annuus	illumina	17.07	Non-normalized	California	-119.41	36.72
Manteca4	annuus	illumina	16.2	normalized	California	-121.23	37.78
Manteca8	annuus	illumina	16.73	normalized	California	-121.23	37.78
14TB-2	annuus	illumina	16.87	Non-normalized	Texas	-98.63	28.97
2OTB-7	annuus	illumina	18.16	Non-normalized	Texas	-98.58	28.18
arg14B-14	annuus	illumina	17.88	Non-normalized	Texas	-97.39	28.70
btm11-1	annuus	illumina	12.29	Non-normalized	Texas	-97.08	27.99
btm3-2	annuus	illumina	17.43	Non-normalized	Texas	-98.27	27.55
btm35-4	annuus	illumina	18.19	Non-normalized	Texas	-97.96	29.53
btm6-1	annuus	illumina	19.43	Non-normalized	Texas	-98.05	27.23
K111	annuus	illumina	24.61	Non-normalized	Texas	-99.14	31.58
TEX	annuus	illumina	12.45	Non-normalized	Texas	-99.14	31.58
SAW3	annuus	illumina	5.5	normalized	Australia	-33.29	137.47
CAW	annuus	454	0.35	normalized	California	-119.00	36.00
INW	annuus	454	0.49	normalized	Indiana	-86.15	40.23
KAW	annuus	454	0.6	normalized	Kansas	-98.60	38.34
UTW	annuus	454	0.53	normalized	Utah	-111.71	39.30
HTAI	annuus	454	0.48	normalized	Texas	-99.14	31.58
CAN.454Reads	annuus	454	0.45	normalized	California	-119.00	36.00
EUW1	annuus	454	0.63	normalized	Europe	-	-
EUW2	annuus	454	0.67	normalized	Europe	-	-
ISI	annuus	454	0.33	normalized	Israel	-	-
QLD	annuus	454	0.52	normalized	Australia	-	-
WAN	annuus	454	0.91	normalized	Australia	-	-
CON2	annuus	454	0.52	normalized	Colorado	-105.78	39.55
IOW	annuus	454	0.99	normalized	Iowa	-93.08	41.87
KSN	annuus	454	0.59	normalized	Kansas	-98.48	39.00
MOW	annuus	454	0.46	normalized	Missouri	-92.17	36.92
NDW	annuus	454	0.56	normalized	North Dakota	-102.79	46.88
NMN	annuus	454	0.45	normalized	New Mexico	-105.02	34.97
OKW	annuus	454	0.56	normalized	Oklahoma	-97.09	35.44

UTN1	annuus	454	0.48	normalized	Utah	-111.08	39.32
arg1820.white	argophyllus	454	0.46	normalized	Texas	-98.13	26.89
HDW	debilis	454	0.47	normalized	Texas	-101.00	32.00
PL109	petiolaris	454	1	normalized	New Mexico	-104.63	34.86

Supplementary Table S2 | Proportion of amino-acid substitutions driven by positive selection.

Different criteria for "fixed sites" (F_{ST} = 1, top one or top five percentile of F_{ST} distribution) were used to calculate *alpha*.

	Comparisons			
	H. annuus* - *H. petiolaris	***H. annuus* - *H. debilis***	***H. debilis* - *H. argophyllus***	***H. petiolaris* - *H. argophyllus***
alpha (F_{ST} = 1 criterion for fixed sites)	0.47	0.43	0.20	0.19
alpha (top 1% criterion for fixed sites)	0.41	0.44	0.20	0.19
alpha (top 5% criterion for fixed sites)	0.42	0.40	0.14	0.14

Supplementary Table S3 | Expected number of islands. Number of islands expected if divergent markers were randomly distributed throughout the genome. To produce random distributions, we permuted all F_{ST} values (resampling, without replacement) for mapped markers, ten times, for each of the four comparisons separately. Sizes and numbers are averaged based on ten permutations and did not differ between comparisons (one-way ANOVA: $F_{3,36} = 0.25$, p-value = 0.8 for islands number; $F_{3,36} = 2.38$, p-value = 0.09 for islands size).

	Comparisons			
	H. annuus* - *H. petiolaris	***H. annuus* - *H. debilis***	***H. debilis* - *H. argophyllus***	***H. petiolaris* - *H. argophyllus***
Number of islands	8	7	8	8
Mean size of islands (cM)	0.00	0.02	0.03	0.01

Supplementary Table S4 | Co-expression clusters and F_{ST} islands. Proportion of loci that occur in both co-expression clusters and F_{ST} islands.

	Comparisons			
	H. annuus* - *H. petiolaris	***H. annuus* - *H. debilis***	***H. debilis* - *H. argophyllus***	***H. petiolaris* - *H. argophyllus***
Number of loci	24,406	24,406	24,406	24,406
Percentage of loci co-occurring in both clusters and F_{ST} islands				
Observed	2.06	2.06	2.17	2.18
Expected	1.63	1.53	1.71	1.66
Chi-square test				
χ^2	26.2	42.61	28.7	39.0
p-value	<< 0.001	<< 0.001	<< 0.001	<< 0.001
df	1	1	1	1
Regression of F_{ST} on coefficient of coexpression				
r^2	0.0034	0.0037	0.0013	0.0006
p-value	<< 0.001	<< 0.001	<< 0.001	0.02

Supplementary methods

Construction of the reference transcriptome

The mRNAseq approach described in the main text was also used to develop a reference transcriptome for a cultivated sunflower line, HA412-HO, which is also the target of genome sequencing efforts[23]. Briefly, one lane of Illumina GAII sequence (2x100 bp) was generated for each of the following four treatment or tissue libraries: leaves of plants exposed to moderate cold (48h at 4°C) or drought (water withheld for circa 48h until wilting) stress, pooled flower heads and leaves, and pooled roots and stems from plants grown under normal conditions.

For construction of the reference transcriptome, sequences from each of the four lanes were cleaned to remove low quality reads and potential contaminating vector sequences with SNOWHITE v.1.1.4[54] and then all data were concatenated and assembled using TRINITY[55] with default parameters. Following this, the longest transcript of each gene (in case of alternative splice variants) over 400 base pairs long was selected for the final assembly (mean length: 998 base pairs), yielding a reference set of 51,468 contiguous expressed sequences (contigs).

Genetic map

Our core mapping population was derived from two highly homozygous sunflower cultivars: RHA280, a confectionary line, and RHA801, an oil seed line. Eighth generation recombinant inbred lines (RILs) of single seed descent were used for mapping. Whole genome shotgun (WGS) sequencing was carried out on the Illumina HiSeq at the McGill University and Genome Quebec Innovation Center (2 X 100 bp). One lane of sequence (~10x genome coverage) was generated for each parent. Eight lanes were then multiplexed with 12 RILs per lane, producing about 1X of coverage for each barcoded sample.

Parental reads were assigned to our draft reference assembly using BWA[41]. Genotypes were called using SAMTOOLS (MPILEUP)[42]. In each individual, genomic contigs were called as descended from one or the other parent based on the presence of at least nine SNPs, with at least 90% called as descended from one or the other parent. There is a trade-off between stringency and marker density and based on preliminary analyses, we found that nine SNPs eliminated mis-scoring while maximizing the number of contigs that could be mapped. An initial set of 4,274 contigs were ordered with MSTMAP[24]. MSTMAP groups markers based on the minimum sum of recombination events (Hamming distance) between their segregation patterns and divides them into linkage groups if the sum is significantly different than observed across all markers. Markers on each linkage group are then ordered using a recursive minimum spanning tree algorithm. Kosambi's mapping function[56] was used to calculate the map distance between adjacent pairs of markers ordered by MSTMAP. The remaining contigs that contained segregating SNPs were added to this initial template map by comparing segregation patterns with markers on the template map in both forward and reverse order. Contigs with an exact match to template markers were binned; others were placed in the most likely position between the best forward and reverse match. In all, 261,999 contigs were placed on the map (mean contig length = 2,417 bp, N50 = 3,517 bp).

To validate the new sequence-based map, markers from a previously published 10,800 locus genetic map[53] developed using an Illumina SNP genotyping array were matched (BLASTN[57]) to the contigs in our new map, and synteny between the two maps was compared. Of 4,727 shared markers, 90.4% are in 17 syntenic blocks as expected and marker ordering is highly conserved (Supplementary Fig. S3). The roughly 10% of non syntenic hits can be explained by picking the second best BLASTN hit if the true homolog is fragmented into several contigs, or if the sequence is present in multi copy.

Co-expression of neighbouring genes

To test whether neighbouring genes tended to have correlated patterns of expression, we used an approach developed in previous studies of expression correlation[27,58,59]. Using only the *Helianthus* accessions from non-normalized transcriptome libraries (Supplementary Table S1), we calculated Pearson correlations for the transcript read counts for all possible pairwise combinations of genes on each chromosome. We then used a sliding window analysis with a window size of eleven genes (five genes on either side of the focal locus) to calculate the average Pearson correlation coefficient within the window. To evaluate the significance of these measurements, we permuted the order of the genes on chromosomes and repeated the sliding-window analysis, performing 1,000 permutations to generate a null distribution. Any loci with average correlations above the 95th percentile of the permuted distribution were considered to be significant co-expression clusters. While it is not possible to identify which of these clusters are statistical artefacts, 10.9% of all loci were identified as co-expression clusters, which is substantially higher than the 5% of loci that would be expected given type I error rates (See Supplementary Fig. S7).

To test whether the co-expression clusters that we identified tended to occur in the same regions of the chromosomes as the islands of divergence, we used χ^2 goodness-of-fit tests of the observed rate at which loci co-occur in both co-expression clusters and F_{ST} islands, compared to the expected rate of co-occurrence if the individual probabilities of occurrence were independent ($\Pr[coexpression\ cluster \cap F_{ST}\ island] = \Pr[coexpression\ cluster] * \Pr[F_{ST}\ island]$). We also note that the χ^2 tests will overestimate statistical significance, because they do not account for auto-correlation in state between adjacent loci (neighbouring loci are more likely to have the same state).

Supplementary references

53. Bowers, J. E. *et al.* Development of a 10,000 Locus Genetic Map of the Sunflower Genome Based on Multiple Crosses. *G3: Genes | Genomes | Genetics* **2**, 721–729 (2012).
54. Barker *et al.* EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evol. Bioinform* **6**, 143–149 (2010).
55. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol* **29**, 644–U130 (2011).
56. Kosambi, D. D. The estimation of map distances from recombination values. *Ann. Eugenic* **12**, 172–175 (1943).
57. Altschul, S. F., Gish, W., Miller, W. & Myers, E. W. Basic local alignment search tool. *J mol. ecol* **215**, 403–410 (1990).
58. Zhan, S., Horrocks, J. & Lukens, L. N. Islands of co-expressed neighbouring genes in *Arabidopsis thaliana* suggest higher-order chromosome domains. *Plant J* **45**, 347–357 (2006).
59. Schmid, M., Davison, T. S., Henz, S. R. & Pape, U. J. A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet* **37**, 501–506 (2005).