www.sciencemag.org/content/344/6185/738/suppl/DC1

Supplementary Materials for

Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation

Víctor Soria-Carrasco, Zachariah Gompert, Aaron A. Comeault, Timothy E. Farkas, Thomas L. Parchman, J. Spencer Johnston, C. Alex Buerkle, Jeffrey L. Feder, Jens Bast, Tanja Schwander, Scott P. Egan, Bernard J. Crespi, Patrik Nosil*

*Corresponding author. E-mail: p.nosil@sheffield.ac.uk

Published 16 May 2014, *Science* **344**, 738 (2014)
DOI: 10.1126/science.1252136

This PDF file includes:

Materials and Methods

Flow cytometric genome size estimation.

Flow cytometry was used to determine genome size as described in (*26*). In short, neural tissue dissected from individual heads of male (n=2) or female (n=2) adult *T. cristinae* was placed into 1ml of Galbraith buffer along with the head of a female *Drosophila virilis* standard (1C = 328 Mb). The sample and standard were co-ground with 15 strokes of the "B" pestle in a 2ml Dounce tissue grinder (Kontes USA). The resultant solution was filtered through a 20-micron nylon mesh, stained with 0.25 mg of Propidium Iodide and adjusted to a final volume of 1ml with Galbraith buffer. The stained solution was mixed using a vortex and then held in the cold and dark for a minimum of 30 minutes. To score genome size, mean PI fluorescence of the stained co-prepared nuclei from the sample and standard was scored using a FACScan (Beckman USA) flow cytometer with excitation at 488 nm. Red fluorescence was scored using a 590nm long pass filter. The red PI fluorescence peak produced by the 2C nuclei of the standard was set to channel 200. To ensure that only nuclei free of cellular tags were used in the assay, a gate was set on side scatter; only nuclei with uniformly low side scatter (small uniform size) were scored for fluorescence. A minimum of 1000 gated diploid nuclei were scored for the standard and for the sample. The coefficient of variation was < 3.0 for all scored peaks. To determine the amount of DNA in the sample, the mean channel number of the peak produced by the diploid nuclei of the standard and the sample were scored using FacScan software. The total DNA amount in the sample was determined as the ratio: average channel number of the sample 2N/ average channel number of the standard 2N times the 1C amount of DNA in the standard. Detailed results were as follows: females, 1C = 1381.8 Mb +/- 3.1 Mb; males, 1C = 1273.8 +/- 0.9 Mb.

De novo assembly.

The first version of the *T. cristinae* genome was presented in (*18*). We summarize here the steps taken in the previous study (see original study for further details). We constructed one fragment library based on 170 bp fragments, and jumping libraries with insert sizes of 500, 800, 2000 (two replicates) and 5000 bp for paired-end sequencing on seven lanes of the Illumina HiSeq 2000 platform with V3 reagents. Genomic libraries were constructed from genomic DNA that was isolated from a total of 10 individual female *T. cristinae* using Qiagen DNeasy Blood and Tissue kits (Qiagen, Inc. USA). To minimize genetic differences that might affect assembly of sequences all individuals used in DNA extractions were collected from the same sample population during spring 2011 (population code: PC). All library construction and sequencing was carried out at BGI, Hong Kong. Sequencing was accomplished on seven lanes (one lane each for all libraries except the 5000 bp library, which was sequenced in two lanes). Raw sequencing reads totaled 303.073 Gigabases (Gb) and following initial quality control sequencing reads totaled 182.015 Gb. Initial quality control included removing reads with greater than 5% N's or with evidence of polyA regions, removing reads where 20% or more of the calls were considered low quality bases, removing adaptor polluted reads, removing reads with overlapping reads, and removing duplicated reads. Three assemblies with different choices of k-mer-size (31, 47, and 71) and the software SOAPdenovo (version 1.05)(*27*) resulted in many small scaffolds that were evidently under-assembled relative to expected genome size. This may be due to heterozygosity within and among the outbred individuals in the sequencing libraries. In contrast, assembly of the reads from the 170, 2000, and 5000 bp insert libraries with ALLPATHS-LG (version 43375)(*28*) yielded a much smaller number of larger scaffolds (the 500 and 800 bp libraries cannot be used in ALLPATHS-LG). Invoking the HAPLOIDIFY=T option in ALLPATHS-LG, to better utilize reads from heterozygous individuals, yielded a longer assembly with fewer scaffolds than without this option. We used this latter assembly for all further analyses. This assembly included 190,773 contigs in 14,221 scaffolds with the N50 (i.e., the smallest scaffold above which 50% of an assembly would be represented) for the scaffolds being 312,000 bp (with gaps). The estimated fraction (total length of assembly in bp / estimated genome size from flow through cytometry) of the genome represented in this assembly is ~80% (1027063217 / (1.3*10^9) = 0.7900486). The fraction of RNA transcripts discovered from a thorough transcriptome sequencing study (*29*) that are represented in the assembly is 0.95, indicating the overwhelming majority of transcripts are represented. This version of the assembly was used for analyses in (*18*), but did not yet contain any linkage group information or any functional, gene, structural or TE annotation (see below). We updated the genome here to include all these components.

Genome annotation.

We performed structural gene annotation with MAKER 2.28b (*30, 31*), an annotation pipeline specifically designed for *de novo* annotation when little data is available on gene models. The workflow of the pipeline involves: (1) identifying and masking out repeat elements, (2) producing *ab initio* gene predictions, (3) aligning protein and RNA evidence, (4) polishing evidence alignments to identify intron-exon boundaries and splice forms, (5) producing evidence-informed gene predictions, (6) integrating the evidence and synthesizing annotations, and (7) computing evidence-based quality scores and selecting the gene models best supported by evidence. The pipeline was specifically set up as follows. We used RepeatMasker 4.0.1 (*32*) to find and mask repeats with RMBlast 2.2.28 and Tandem Repeats Finder 4.07b (*33*), using RepBase 18.05 (*34*) and Dfam 1.2 (*35*) curated libraries. For *ab initio* gene predictions, we used SNAP 2013-02-16 (*36*) and GeneMark-ES 2.3.e (*37*). SNAP finds protein-coding regions using a hidden Markov model where exon-intron boundaries are modeled as transitions between hidden states. SNAP requires supervised training, and therefore we used CEGMA 2.4.010312 (*38*) to produce an initial training gene model by finding orthologs of a set of highly conserved eukaryotic proteins. CEGMA was set up to use NCBI BLAST+ 2.2.28+ (*39*) to identify candidate genes, and Wise 2.4.1 (*40*), HMMER 3.1b1 (*41*), and geneid 1.4.4 (*42*) to refine gene structures and determine exons and introns. From a subset of 248 ultraconserved genes, we found 135 complete matches (54.4%) and 214 partial matches (86.29%) in the *T. cristinae* genome. Genemark-ES follows a similar hidden Markov model approach, but it does not require supervised training, as it uses a heuristic method to initialize the searching algorithm. We used a dataset of 835,424 proteins of insects retrieved from UniProt Knowledgebase (*43*) (downloaded on 26/06/2013) as protein evidence. Regarding RNA evidence, we used the already published *T. cristinae* transcriptome (*29*), along with the transcriptomes of the two most closely related species available so far: *Phyllium philippinicum* and *Embioptera sp.* (*44*). We retrieved the publicly available 454 data for these species (GenBank SRR399296, SRR399289), pre-processed and cleaned the sequences with SeqTrimNext 2.0.54 (*45*), and assembled them *de novo* using Trinity r2013-02-25 (*46*). Intron-exon boundaries were refined using Exonerate 2.2.0 (*47*). To evaluate the quality of the annotations, MAKER 2 computed Annotation Edit Distances (AED). AED is based on the protein and RNA evidence that supports a given annotation, ranging from zero when the annotation is in total agreement with its evidence, to one when there is no evidence at all supporting the annotation (*48*).

Functional annotation was carried out using InterProScan 5.4.25.0 (*49*). We scanned *T. cristinae* predicted proteins against 11 signature databases: COILS 2.2, Gene3D 3.5.0, PANTHER 8.1, Pfam-A 27.0, PIRSF 2.84, PRINTS 42.0, ProDom 2006.1, PROSITE 20.97, SMART 6.2, SUPERFAMILY 1.3, and TIGRFAMs 13.0. The scan of the 44,292 protein sequences yielded a total of 147,560 matches distributed as follows: 3,808 hits for COILS, 22,294 for Gene3D, 30,096 for PANTHER, 25,474 for Pfam-A, 416 for PIRSF, 9498 for PRINTS, 3 for ProDom, 20,084 for PROSITE, 13,643 for SMART, 21,474 for SUPERFAMILY, and 770 for TIGRFAMs. We found 23,083 predicted proteins that had at least one match with any of the databases. We limited further functional analyses to Pfam-A matches, because this database is characterized by high-quality, manually curated entries. We extracted the Gene Ontology (GO) terms that mapped to the Pfam domains that matched the predicted proteins.

In summary, we annotated a total of 44,292 genes, 188,964 coding DNA sequences (CDS), and 3,556 untranslated regions (UTRs). Genic regions (coding or non-coding) comprised 22.5% of the genome (190,238,410 bp) and 16.9% (32,233,651 bp) of these genic regions were coding regions. ‘Genic non-coding’ regions were defined as introns and UTRs. The mean AED was 0.38 (95% interval = 0.05 - 0.85), and 33,482 genes had an AED below 0.5 (75.6%). Regarding functional annotation, we found 23,083 predicted proteins that had at least one match with any of the functional databases used. From these proteins with functional predictions, 16,307 matched 25,474 Pfam domains associated with at least one Gene Ontology (GO) term. We found 439 unique GO terms for biological process, 156 for cellular component, and 568 for molecular function.

Transposable element (TE) annotation.

De novo repeat detection was done using raw reads and assembled scaffolds. Consensus sequences for repeats with at least 10 copies and minimum 200 nucleotides (nt) in length were computed with ReAS and RepeatScout (*50, 51*). Output sequences were pulled together and redundant repeats were clustered with 95% identity threshold using UCLUST with the centroid option (*52*). In a first round of automatic classification with REPCLASS (*53*), using homology to RepBase (*54*) entries and structural features, repeats at least annotated to TE superclass level (class I or class II) were kept in a putative TE library, whereas non-TE entries were discarded. Boundaries of elements longer than 1000 nt or elements classified to super-family level (e.g., hAT, gypsy, etc.) were manually extended where possible. This was done by blasting each repeat against the whole genome (evalue cutoff: 1e^{-20}), aligning best hits plus extended boundaries with Muscle (*52*) and curating alignments by eye. Repeats were considered ‘complete’, if boundaries could not be extended any further with at least three sequences aligned. Consensus repeats, representing the ancestral repeat state, were generated with these alignments following the majority rule. In a second round of classification, extended elements were annotated with REPCLASS (as above) and with RepeatMasker (*32*), tblastx and blastn against RepBase and non-redundant NCBI entries (keywords: retrotransposon, transposase, reverse transcriptase, transposon, transposable element)(e-value > 1e^{-30}). Ambiguous annotations were re-checked with Censor (*54*), and by manual curation using knowledge of TE family structure. The final 'strict' (no unknown or ambiguous repeats) species-specific *T. cristinae* TE library contained 904 reconstructed repeats with an average length of 1470 nt. In summary, via this gene, functional, and TE annotation we identified 16,307 genes with predicted function and the proportion of the assembly in different categories was 4% coding, 19% genic non-coding, and 77% intergenic, with a total TE content of 24%.

Linkage mapping.

We estimated recombination rates between SNPs on different scaffolds and used this information to delineate linkage groups and order DNA sequence scaffolds within linkage groups, similar to the approach used with the recent *Heliconius* genome (*9*). We generated three F1 mapping families from wild-caught stick insects from three locations (source population 'M' on *Adenostoma*, 34 30.897 120 04.278; source population 'S' on *Ceanothus*, 34 31.338 119 49.877; source population 'WT' on *Adenostoma*, 34 30.950 120 04.389). Each family included two parents from different populations (female by male cross: S x M, WT x M, and M x S), yielding 114, 48, and 24 sequenced offspring, respectively. We isolated genomic DNA from these stick insects using Qiagen's DNeasy Blood and Tissue kit. We constructed reduced-complexity genomic libraries for genotyping-by-sequencing (GBS) following published protocols (*16*). Briefly, we digested genomic DNA from each individual using two restriction enzymes, EcoRI and MseI. We then used T4 DNA ligase to attach adaptor oligos containing a unique 8, 9, or 10-bp barcode sequence and the Illumina sequencing adaptor to the digested fragments. We PCR amplified these fragments using the Illumina sequencing primers. Following PCR, we pooled all genomic libraries and size selected DNA fragments ranging in size from 250–400. We gel-purified these fragments using QiaQuick gel purification kit (Qiagen, Inc.). These libraries were sequenced over three lanes on the Illumina HiSeq 2000 platform at the National Center for Genome Resources in Santa Fe, NM. We obtained 157,925,171 DNA sequences with a maximum read length of 100 bp.

We used the BWA-backtrack algorithm implemented in bwa 0.7.5a-r405 (*55*) to align these sequences from each individual to the *Timema* genome scaffolds. We discarded bases with quality scores less than 10, allowed a maximum edit distance of 4 between the read and reference sequences, and only placed reads with a unique best match. We used a 20 bp seed with a maximum edit distance of two to increase the speed of the alignment method. We then used the Bayesian variant caller implemented in samtools and bcftools 0.1.19 (*56*) to identify single nucleotide variants from the aligned sequence data. We defined a site as variable if the probability of the data under the null hypothesis (no variation at the site) was less than 0.01 using the full prior with $\Phi = 0.001$. We required data for 90% of individuals to designate a variable locus, and we identified variable loci separately for each mapping family. We identified 79,109 single nucleotide variants using these criteria with an average coverage of 6.2 sequences per SNP per individual. We then extracted the subset of SNPs that were fully recombination informative in each family (i.e., homozygous in one parent and heterozygous in the other), and where we had moderately high confidence in the parent genotypes. Specifically, we calculated the posterior probability of each genotype from the genotype likelihoods with a uniform prior, and only retained the 21,125 SNPs where one parent was homozygous and the other was heterozygous with posterior probabilities of at least 0.8.

We developed and implemented a Bayesian-Monte Carlo method to estimate recombination rates between these SNPs. We first calculated the posterior probability of each offspring genotype as Pr(g | x, s) = (Pr(x | g) Pr(g | s)) / Pr(x), where Pr(x | g) is the probability of the sequence data and quality scores (x) given each genotype (g; this is the genotype likelihood from bcftools), Pr(g | s) is the prior probability of the offspring genotype given parent genotypes and Mendelian segregation (s), and Pr(x) is the marginal likelihood of the data. We then repeatedly (100 times) sampled offspring genotypes according to their posterior probabilities and calculated the observed recombination rate from the bi-locus genotype frequencies. We only included offspring in the recombination rate calculation for a locus pair if the posterior probability of their most likely genotype was at least 0.9. Even with this high stringency, uncertainty in genotype introduces bias in recombination rate estimates such that the recombination rate between tightly linked loci will tend to be overestimated. Thus, whereas these recombination rate estimates allow us to generate a hypothesis regarding the partitioning of scaffolds into linkage groups and ordering of scaffolds within linkage groups, they are not accurate enough to determine the distance between scaffolds.

We developed a heuristic method to designate linkage groups and order scaffolds based on the mean recombination rate between SNPs on each pair of scaffolds. We used all three families and recombination rates estimated in the mother and father of each family to estimate the average recombination rate between each pair of scaffolds. Thus, the linkage map reflects the average recombination rates across six individuals (three families with two parents in each family) rather than the recombination rates in any single individual. We first use principal component analysis (PCA) to statistically summarize the matrix of pairwise mean recombination rates between scaffolds. We retained the first 13 PC axes, which explained 11.2% of the variation, and correspond with the number of linkage groups in *T. cristinae*. We then used K-means clustering with the Hartigan-Wong algorithm to define the center of 16 clusters of scaffolds based on these 13 PC axes. Next, we assigned individual scaffolds to putative linkage groups with linear discriminant analysis based on the same 13 PC axes and cluster means from K-means clustering (*57*). We retained 13 clusters with assigned scaffolds where the mean recombination rate among scaffolds assigned to the cluster was less than 0.3. These analyses were implemented in R. Finally, we employed an R implementation of the unidirectional growth algorithm (*58*) to order scaffolds along each linkage group (the R implementation of this algorithm we used was written by Q. Wang and A. Buerkle). We treat the resulting linkage group designation and ordering as a working hypothesis for the genome organization of *T. cristinae* that will be iteratively improved over time.

Whole genome re-sequence data alignment and variant calling.

We extracted DNA from 160 wild-caught stick insects from eight populations (these are the same individuals examined in a previous genotype-by-sequencing study for a much smaller set of 86,130 SNPs, see (*16*) for details). We then fragmented each individual's DNA to an average size of 150 bp to prepare fragment libraries for DNA sequencing. . We sequenced these libraries on eight lanes of Illumina HiSeq 2000 with V3 reagents. We obtained paired-end reads that were 100 bp long, i.e., 100 bp per read and thus 200 bp per pairs for each fragment. Library preparation and sequencing was conducted at the Welcome Trust for Human Genetics in Oxford, UK. Sequencing yielded a total of 3,166,737,472 reads. We used BWA-MEM algorithm implemented in bwa 0.7.5a-r405 (*55*) to align paired-end whole genome re-sequence data from the 160 individuals to the *T. cristinae* draft genome. We set the minimum seed length to 20 bps, the internal seed tuning parameter to 1.3, and the minimum aligned read quality score to 30 (this parameter specifies the minimum phred-scaled probability that the read is not correctly mapped). We then used samtools and bcftools 0.1.19 (*56*) to sort and index the alignments and identify variable nucleotides. We required sequence data for 124 of the 160 sequenced stick insects to designate a variable locus and we defined a site as variable if the probability of the data under the null hypothesis was less than 0.001, using the full prior with theta equal 0.001. We found 12,287,179 single nucleotide variants (median and mean coverage per SNP per individual = 2.8 and 5.0, respectively; average SNP density was 8.4 SNPs per 1000 bp). These included 4,391,556 SNPs with a minor allele frequency greater than 1% that mapped to one of the 13 putative linkage groups. The following analyses used these 4,391,556 SNPs.

Genome re-sequence data from additional *Timema* species.

To construct a rooted phylogenetic tree of the *T. cristinae* populations (see below) we generated whole genome re-sequence data from five outgroup species of *Timema* that are closely related to *T. cristinae* (one individual each of *T. poppensis, T. californicum, T. petita,* and *T. landelsensis,* and two specimens of *T. knulli*), using the same method as described for *T. cristinae.*

Phylogenetics.

We carried out phylogenetic analyses using the previously defined 4,391,556 SNPs for 160 *T. cristinae* individuals. We used a custom Perl script to generate a multiple alignment from the genotype with the highest likelihood and coding heterozygotes as IUPAC ambiguities. We estimated the maximum-likelihood phylogenetic tree using ExaML 1.0.9 (*59*), which is an implementation of the RAxML search algorithm (*60*) optimized for improved parallel efficiency on computer clusters. We first used RapidNJ 2.3.0.2 (*61*) to infer a Neighbor-Joining tree under a Kimura's substitution model. This tree was used as a starting tree for ExaML to conduct maximum-likelihood inferences under a GTR + Γ substitution model. Due to computational limitations, approximately-maximum-likelihood bootstrap tree inferences were computed with FastTree 2.1.7 (*62*), using the ExaML maximum-likelihood tree as starting tree and the GTR + CAT approximation model with 20 rate categories (100 bootstrap replicates). This tree showed strong grouping of individuals by geography, not host (Fig. S2). However, four individuals from two geographically distant populations grouped together. These individuals presented long branches and high mean neighbor-joining genetic distances when compared to the rest of the individuals (Fig. S2), suggesting possible long-branch attraction artifact (LBA)(*63*). Likewise, analyses including the other five *Timema* species resulted in strong clustering of the same four *T. cristinae* individuals with the other species, adding further evidence of LBA (aberrant individuals were not evident in any non-phylogenetic analyses such as principal components analyses, see below). Thus, we excluded these four individuals from subsequent phylogenetic analyses to avoid artifacts due to LBA. We generated this new dataset by using samtools and bcftools to identify variable positions as before, but also considering the alignments of the other five species of *Timema*, which were used as the outgroup. From the 56,704,814 SNPs found, we retained those where at least 90% of the individuals were sampled for subsequent phylogenetic analyses. As before, we used a custom Perl script to generate a multiple alignment that included 162 individuals (156 *T. cristinae* individuals and 6 individuals from the other five *Timema* species) and 3,199,186 variable positions. The maximum-likelihood tree and bootstrap support were estimated as before. Additionally, we specifically tested whether *T. cristinae* individuals grouped by host plant. We used ExaML to reconstruct trees as before, but constraining tree searches to group individuals by host. Next we calculated site-wise likelihoods with RAxML 8.0.3 (*64*), which were subsequently used with consel 0.20 (*65*) to perform an approximately unbiased (AU) test (*66*). Grouping by host plant was significantly rejected regarding of using the initial 160 *T. cristinae* dataset ($p = 1 \times 10^{-42}$) or the final dataset of 162 individuals that included the outgroup species ($p = 4 \times 10^{-56}$).

Principal Components Analysis.

We used principal component analysis (PCA) to statistically summarize the distribution of genomic variation across the 160 wild-caught *T. cristinae*. We first calculated the Bayesian posterior mean genotype for each individual and locus as Σg = {0,1,2} g * Pr(x | g) * Pr(g), where Pr (x | g) is the genotype likelihood calculated with samtools and bcftools, and Pr(g) is the prior genotype probability (we specified an uninformative prior probability of 1/3 for each genotype). We then calculated the N x N genetic covariance matrix based on the genotype estimates and transformed this matrix with PCA. We computed genetic covariance matrixes and performed PCA in R (*67*) using the prcomp function. The first PC explained 66.6% of the variation in genomic similarity among individuals. We fit a linear random effects model for the PC1 scores to estimate the proportion of genomic variation explained by host plant and locality with the lmer function in the R package ade4 (*68*).

Population differentiation (F_{ST}).

We quantified genome-wide genetic differentiation between the four pairs of *Adenostoma* and *Ceanothus*-feeding stick insect populations. Specifically, we estimated Hudson's F_{ST} (*69,70*) for each SNP and population pair as F_{ST} = 1 - Hw/Hb = 1 -(p1 * (1-p1) + p2 * (1-p2))/(p1 * (1-p2) + p2 * (1-p1)). Here Hw is the mean number of differences between sequences sampled from the same population, Hb is the mean number of differences between sequences from different populations, and p1 and p2 are the reference allele frequencies in population 1 and 2. We estimated the sample allele frequencies directly from the sequence data and quality scores as p = 1/2n * sum_n sum_g = {0,1,2} g * Pr (x | g), where Pr (x | g) is the genotype likelihood calculated with samtools and bcftools (*71*). We conducted this analysis in R with code written by the authors.

Quantifying genomic differentiation using a Hidden Markov model approach.

We used a discrete state, homogeneous Hidden Markov model (HMM) to delineate contiguous regions of the genome with different levels of genetic differentiation between the *T. cristinae* population pairs. This allowed us to examine the number, size, and distribution of regions of divergence. A similar approach was used by Hofer et al. (*22*) to statistically summarize genetic differentiation among human populations. We assumed that the genome could be delineated into contiguous genetic regions characterized by low, intermediate, or high genetic differentiation (F_{ST}). These classes of loci are hidden states in the HMM and our main goal was to infer these hidden states. We modeled logit(F_{ST}), and assumed that the distribution of logit(F_{ST}) varied among states. Specifically, we assumed a Gaussian distribution of logit(F_{ST}) for each state with the variance fixed at the genome-wide sample variance. We estimated the mean logit(F_{ST}) for each state and the transition rate matrix among states from the data using the Baum-Welch algorithm (*72*). Following (*22*), we disallowed transitions directly between low and high genetic differentiation states. We then used the Viterbi algorithm to predict the most likely sequence of hidden states from the data and estimated parameters. We used the R package HiddenMarkov (*73*) for these analyses. We were primarily interested in the regions of high divergence from the HMM analyses, and whether they correspond to regions affected by selection in the field experiment (see below). However, the model also delimited a number of regions of low divergence, which tended to be small. Further work is required to determine the significance of these regions of low divergence, which for example, could be subject to purifying selection.

Quantifying parallelism of individual SNPs.

We compared genetic differentiation (F_{ST}) for individual SNPs across population pairs to determine the extent to which divergence between *Adenostoma* and *Ceanothus*-feeding populations consistently involved the same SNPs (i.e., parallelism). We focus these analyses of parallelism on SNPs to increase precision, but note that SNP-based and HMM-model based results were strongly associated and highly congruent (details below). We first identified the set of SNPs above the 90th empirical F_{ST} quantile between each population pair (i.e., the top 10% of F_{ST} for each population pair). We then enumerated the number of population pairs where each SNP was in this top quantile set. We then repeatedly (1000 times) randomized SNP labels among population pairs to generate a null distribution for the number of SNPs that would be expected to be in the top 10% for more than one population pair by chance. We generated this null distribution considering SNPs that were in the top quantile set in two, three, or all four population pairs, and for those in two or more population pairs. We contrasted these null distributions with the observed counts. We implemented this analysis in R. We then conducted a similar analysis for numbers of observed versus expected SNPs for the 99th (rather than 90th) quantile of the empirical F_{ST} distribution (1000 randomizations).

Correspondence between divergence and parallelism of SNPs and HMM regions.

We tested whether parallel high divergence SNPs (those in the top 10% of the empirical F_{ST} distribution for two or more population pairs) occurred in parallel high divergence genetic regions delimited by the HMM. Specifically, we tested whether parallel high divergence SNPs were classified as having the high divergence HMM state in two or more populations pairs more often than expected by chance. To achieve this, we first computed the empirical 5x5 contingency table for the number of populations pairs (0, 1, 2, 3, or 4) where a SNP exhibited high divergence based on its estimate of F_{ST} versus the number of populations pairs where a SNP exhibited high divergence based on its HMM state. We then randomized locus ids and recalculated the contingency table 1000 times to generate null expectations for the correspondence between divergence and parallelism of SNPs based on F_{ST} and HMM states. We then quantified the extent and statistical significance of excess correspondence beyond these null expectations. Results from these analyses are presented in Table S2 and Figure S5 and show that parallel high divergence SNPs occur much more often in parallel high divergence HMM regions (i.e., these SNPs were assigned high divergence HMM states in multiple populations) than expected by chance.

Executing the field experiment.

During a one-week period in March 2010, *T. cristinae* (n = 2350) were collected from a phenotypically and genetically variable population (LA) of the host *Adenostoma fasciculatum*. This population is genetically variable but within the range of other populations (Fig. S3). A random sample of 300+ of these insects collected from throughout the host-plant patch was kept in 90% ethanol as representative ‘ancestors’ of the experimental populations (31 of these were used for sequencing, see below). The remaining 2000 *T. cristinae* were kept alive, in Nalgene bottles in groups of 25 or 50, for transplantation onto the experimental bushes.

The host-shift was conducted at the same site as a previous mark-recapture experiment, but using different individual bushes (*74*) (with the following GPS coordinates: Block 1A N34 30.859 W119 47.986; Block 1C N34 30.859 W119 47.986; Block 2A N34 30.822 W119 47.961; Block 2C N34 30.851 W119 47.969; Block 3A N34 30.862 W119 48.031; Block 3C N34 30.862 W119 48.031; Block 4A N34 30.831 W119 48.112; Block 4C N34 30.828 W119 48.091; Block 5A N34 30.880 W119 48.103; Block 5C N34 30.867 W119 48.113; Fig. S6 for map). Bushes were chosen and prepared for *T. cristinae* to be shifted onto them as follows. The experiment consisted of five paired blocks. Each block consisted of one plant individual of each host species (‘experimental bushes’ hereafter). The experimental bushes were not touching one another and within each block were generally separated from each other by several meters. Different experimental bushes within blocks were closer to each other than to experimental bushes in other blocks. When necessary, other plants near experimental bushes were chopped down using clippers. Prior to experimental introduction of the 2000 insects mentioned above, any *T. cristinae* on the experimental bushes were removed using the following protocols. Insects were stripped from each experimental bush with sweep nets each day from March 20-26th, alternating stripping in the morning versus afternoon, and not returned to the bushes upon which they were captured. Past studies have shown this is a highly effective method for removing *T. cristinae* from a bush (*75*), and indeed by the last day of stripping very few or no *T. cristinae* were captured on our experimental bushes. All non-*Timema* (i.e., other arthropods) captured during the stripping protocol were returned to the experimental bush that they were captured on. At this point, the experimental bushes were ready for *T. cristinae* to be transplanted onto them.

The aforementioned Nalgene bottles harboring the 2000 experimental animals were assigned randomly to block and host. A total of 200 insects were assigned per plant individual. The experimental animals were released onto the experimental bushes by gently shaking the insects from the Nalgene bottles onto the branches and foliage of the experimental bushes. The insects readily clung to the foliage. High densities of the focal species comparable to those used in the experiment exist naturally (*75*), but certainly resource competition could have influenced selection in the experiment. Notably, several previous studies have shown that dispersal by *T. cristinae* across 'bare ground' (grassy regions not containing suitable hosts) is near or even completely absent (*16, 74, 76, 77*). Because our experimental bushes were separated from all other plants by regions of bare ground, there was likely little or no dispersal in our experiment following transplantation. This means that the allele frequency changes we observed (see below) were driven primarily by selection and drift. On March 4th and 5th 2011 (one insect generation later), the experimental bushes were scoured for any *T. cristinae.* In total, 418 bugs were collected. These represent the F1 descendants of the original 2000 founders. A potential advantage to the between-generation analysis employed here (rather than a within-generation one) is that it integrates selection over life history stages and across viability, mating, and fertility selection.

Genotype-by-sequencing of the experiment.

We constructed reduced-complexity genomic libraries for the 418 F1 descendants and 31 'ancestors' as described for the mapping families above. The pooled and size-selected libraries were then sequenced on five lanes of the Illumina HiSeq 2000 system using V3 reagents at National Center for Genome Resources in Santa Fe, NM. This generated 885.3 million 100 bp sequences with identifiable individual barcodes. We used the bwa-backtrack algorithm in bwa version 0.6.2-r126 (aln and samse functions) to align these sequence data to the *T. cristinae* genome assembly. We allowed a maximum of four mismatches between each sequence and the reference genome. We then used the Bayesian model in bcftools version 0.1.17-dev to identify SNPs. We only designated a SNP if reads occurred in a minimum of 95% of the individuals and the probability of the data was less than 0.001 under the null model that all samples were homozygous for the reference allele with the full prior and theta equal to 0.001. We ignored insertions and deletions. This yielded 128,216 SNPs with an average of 9.8 reads per individual per SNP. To better satisfy the assumptions of the method used to quantify allele frequency changes in the experiment (see below), further analyses were restricted to the 82,060 SNPs that were largely uncorrelated with neighboring SNPs. Specifically, we defined every consecutive pair of SNPs on the same scaffold as 'neighboring' and then calculated LD (measured by r^2) between each consecutive pair of SNPs and only retained a SNP for downstream analyses if r^2 between it and the previous SNP was less than 0.05.

Genomic analysis of the experiment (Wright-Fisher model).

We fit a Bayesian model to the sequence data from the experiment to infer allele frequency change and the variance effective population size for each experimental population while incorporating uncertainty in individual genotypes. In particular, we assumed that each experimental population evolved according to the Wright-Fisher model (78-80). In other words we assumed generations were discrete (as is the case in *Timema*) and that the expectation and variance in the allele frequencies in the F1 generation followed a binomial distribution. This required estimating the allele frequencies in the source population and in the F1 generation in each experimental population. We estimated the posterior probability distribution for the source population as $\Pr(\pi, \gamma, \theta \mid y, \varepsilon) \propto \Pr(y \mid \gamma, \varepsilon) \Pr(\gamma \mid \pi) \Pr(\pi \mid \theta) \Pr(\theta)$. The term $\Pr(y \mid \gamma, \varepsilon)$ is the probability of the source populations sequence data (y) given the source population genotypes (γ) and quality scores (ε). We pre-calculate this genotype likelihood using samtools and bcftools following (*74*). $\Pr(\gamma \mid \pi)$ is the probability of the source population genotype data given the source population allele frequencies and Hardy-Weinberg genotype frequency expectations, such that for an individual and locus $\Pr(\gamma = 0) = \pi^2$, $\Pr(\gamma = 1) = 2 * \pi * (1 - \pi)$, and $\Pr(\gamma = 2) = (1 - \pi)^2$. $\Pr(\pi \mid \theta)$ is the conditional prior distribution on the source population allele frequencies such that $\pi \sim \mathrm{Beta}(\theta,\theta)$. Thus, the parameter theta describes the allele frequency distribution.

The source population allele frequencies are a component of the prior distribution for the experimental population allele frequencies (p). Let g_k and p_k denote the genotypes and allele frequencies for each of the ten experimental populations (i.e. k = {1,2,...,9,10}). We model the experimental population sequence data in the same manner as the source population data, specifically we define the genotype likelihood given the data (x_k) and quality scores (e_k) for the experimental population k as $\Pr(x_k \mid g_k, e_k)$ following Li (*72*). We define a beta-binomial prior for the genotypes that reflects the experimental design, such that $\Pr(\text{sum } g_k \mid \alpha, \beta) = (2\ n_k \text{ choose } \Sigma\ g_k)\ B(\Sigma\ g_k + \alpha, 2\ n_k - \Sigma\ g_k + \beta)/B(\alpha, \beta)$. Here $\alpha = \pi * Ne_k$ and $\beta = (1 - \pi)\ Ne_k$, Ne_k is the single generation variance effective population size in population k, and n_k is the number of F1 individuals captured in population k. Thus we model each experimental population as a Wright-Fisher population with Ne_k individuals from the source population contributing to the F1 generation, and we assume that all (or approximately all) n_k individuals were sampled from the F1 generation. A similar model has been adopted to describe population differentiation by genetic drift in natural populations descended from a common ancestral population (*78-80*).

We implemented an MCMC algorithm to obtain samples from the posterior probability distribution of this model. We wrote the computer-based implementation of this model using the C++ programming language and the Gnu Scientific Library (*81*). In addition to the model parameters g, p, π, and Ne, we also generated samples of a derived parameter describing the allele frequency change $dp_k = p_k - \pi$. Importantly the posterior probability distribution for this parameter incorporates uncertainty in the source and experimental population allele frequencies. We ran three independent MCMC analyses with 15,000 iterations, a 5000 iteration burn-in, and a thinning interval of 10. We evaluated convergence to the posterior distribution graphically and quantitatively, and then combined the MCMC samples from the three chains for posterior inference.

We report the minor allele frequency distribution in the source population (π) in Figure S3 (i.e., genetic variation in the 31 putative ancestors). Most of the SNPs we analyzed had relatively high minor allele frequencies in the source population (e.g., $\pi >$ 5%) and thus substantial potential to show a rapid evolutionary response. Whereas there was some uncertainty in the source population allele frequencies, probability theory and our empirical results indicate that the 62 gene copies we analyzed (31 diploid individuals) provided adequate estimates of these allele frequencies (Figure S3).

We summarized the analysis of the experiment by calculating the mean allele frequency divergence between hosts across the five, blocked population pairs for each SNP. This was calculated as $\mathrm{Dp} = \overline{\mathrm{p}A} - \overline{\mathrm{p}C}$. Here $\overline{\mathrm{p}A}$ and $\overline{\mathrm{p}C}$ are the average allele frequency change at a locus across the five experimental populations on *Adenostoma* or *Ceanothus* relative to the ancestral allele frequencies. Thus the parameter Dp measure how much populations transplanted to *Adenostoma* and *Ceanothus* diverged in allele frequencies from each other and relative to the founder population. The expectation is that parallel divergent selection between hosts would drive consistent allele frequency divergence between hosts at a locus across populations.

We next asked whether the SNPs with the greatest parallel allele frequency divergence between hosts during the experiment (specifically SNPs in the top 99.5th empirical quantile of such change) mapped to parallel genetic divergence regions in the natural population pairs based on the HMM. When SNPs in the aforementioned top 99.5th empirical quantile from the experiment were present in both data sets (i.e., experimental and natural populations) we simply assigned the experimental SNP the HMM state of the actual SNP. Otherwise, we assigned the experimental SNP the HMM state of the nearest upstream and downstream SNP, but only if these states matched each other (if the HMM states did not match each other, the SNP was not considered further). We were thus able to assign HMM states of low, moderate, or high divergence to 213 SNPs that were in the top 99.5th empirical quantile for parallel allele frequency divergence between hosts during the experiment.

We then used a permutation test to ask whether these 213 SNPs were assigned a high HMM state between multiple (two or more) natural population pairs more often than expected by chance. We did this by repeatedly (1000 times) sampling sets of 213 SNPs to generate a null distribution for parallel high divergence for random sets of SNPs. The reason we used this method of relating SNPs with the largest divergence between hosts in the experiment to parallel HMM divergence regions in nature (rather than to the F_{ST} value in nature of the specific SNPs themselves) is that the SNPs analysed in the experiment were obtained from one of the eight natural populations we examined (the population 'LA' used to found the experiment) and were generally not variant in all of the other seven natural populations. This means that testing if SNPs showing the strongest experimental responses lie in the parallel HMM divergence regions is the most powerful approach as it considers all the SNPs analysed in the experiment (rather than the subset which are variable in all populations). Nonetheless, we checked the robustness of our result by redoing the analysis described directly above using the subset of SNPs that were variant in both the experiment and in each of the natural populations (n = 3654 SNPs). We found the same result: a greater number of 99.5th quantile parallel divergent change SNPs in the experiment have high HMM divergent states in multiple natural population pairs than expected by chance (observed 7, expected 3, s.d. = 1.6, $p < 0.05$, randomization test).

Tests for enrichment of coding regions and specific functions.

We defined a metric of parallel genetic differentiation that we used to then target SNPs for tests of functional enrichment. To calculate this metric we first determined the empirical F_{ST} quantile of each SNP in each population pair. We then summed these quantiles across the four population pairs. This aspect of the metric is informative regarding consistent high genetic differentiation between population pairs (and is the F_{ST} metric in Table S3), but does not provide information about whether the same allele is consistently at higher frequency in *Adenostoma* vs. *Ceanothus*-feeding natural populations. Thus, we combined this with a quantile-based measure of allele frequency differences by first calculating the difference in the reference allele frequency between each *Adenostoma* and *Ceanothus*-feeding population pair. We then converted these differences to empirical quantiles, and summed them across the four population pairs. We folded these summed quantiles such that consistent negative or positive allele frequencies would be equivalent (which is the allele frequency difference, i.e., AFD, metric in Table S3). We then combined (by simply adding) these two (i.e., F_{ST}-based and AFD-based) summed quantiles for each locus into a composite quantile score (which comprises the 'combined' metric in Table S3 and which is reported in the main text). This score can range from zero to eight, where an eight would mean that the SNP was the most differentiated between all population pairs and had the highest, consistent allele frequency difference between all pairs. Notably, we obtained highly congruent results using each of the three (F_{ST}, AFD, combined) metrics of parallelism (Table S3).

From all the 4,391,556 SNPs, we found that 1,202,332 were in genic regions and 351,370 were in coding regions. We assessed multiple top parallel divergent SNP datasets under various metrics of parallelism and quantile cut-offs (Table S3). In these datasets, the proportion of top parallel divergent SNPs in genic regions (from 20.5% to 38.6%) fluctuated around the proportion found for all SNPs (27.3%). In contrast, the proportion of top parallel divergent SNPs found in coding regions (from 9.9% to 13.6%) was consistently above the proportion for all SNPs (8.0%). We used a randomization approach to test for enrichment in coding regions in the most parallel divergent SNPs. In order to maximize statistical power, the statistical test reported in the main text concerned the most inclusive dataset (the dataset using the combined metric and defined by the 0.0001 quantile cut-off). We used R to estimate an empirical null distribution for the proportion of SNPs in coding regions by randomly drawing 100,000 samples from the total distribution of SNPs. We subsequently computed the empirical cumulative distribution and calculated the 2-tail p-value. The p-values for the different quantiles for the combined metric varied predictably according to sample size and thus statistical power (i.e., according to the number of SNPs in the top set, quantile = 0.00001 contains 44 SNPs, $p = 0.30$; quantile 0.000025 contains 110 SNPs, $p = 0.14$; quantile 0.00005 contains 220 SNP, $p = 0.03$, quantile 0.0001 contains 439 SNPs, $p = 0.0008$).

A similar randomization approach was followed to test for functional enrichment. First, for each one of the SNPs, we retrieved the closest predicted gene within the same scaffold, and collected the GO terms that mapped to the Pfam domains that matched that gene. It is worth noting here that not all SNPs were in scaffolds harboring at least one gene prediction, nor were all predicted genes matched by at least one Pfam domain, nor were all Pfam motifs mapped to at least one GO term. We collected 992,768 matches to 797 unique GO terms associated with 344,433 (7.85%) SNPs. In particular, we found 499,532 matches to 394 unique molecular function terms, 148,803 matches to 106 unique cellular component terms, and 344,433 matches to 297 unique biological process terms. We tested for functional enrichment using the most inclusive of the top parallel divergent SNPs datasets (quantile 0.0001 for the combined metric, see above). In this dataset, we found 122 matches to 74 unique GO terms associated with 49 SNPs. In particular, we found 122 matches to 33 unique molecular function terms associated with 41 SNPs (the results reported in the main text refer to these SNPs in particular), 11 matches to 6 unique cellular component terms associated with 9 SNPs, and 49 matches to 27 unique biological process terms associated with 28 SNPs (Table S4). As before, we used R to calculate an empirical null distribution for the proportion of SNPs associated with each one of the unique GO terms. Due to low statistical power, we excluded the unique GO terms where the number of associated SNPs was ≤ 2. Given the number of tests involved, we calculated Bonferroni-adjusted 2-tail p-values to avoid spurious significance.

In order to further evaluate particular functions involved in parallel divergence, we also examined the molecular function GO terms associated with the 32 HMM divergence regions that contained the SNPs showing the strongest parallel and divergent allele frequency change between hosts in the transplant experiment. We retrieved all genes that overlapped these regions, and collected the molecular function GO terms that mapped to the Pfam domains that matched those genes. We repeated this analysis for each pair of natural populations. In total, we found 47 matches to 16 unique molecular function GO terms. In particular, we found 11 matches to 9 unique GO terms associated with 8 genes for HVA × HVC, 14 matches to 11 unique GO terms for 8 genes for MR1 × MR1C, 14 matches to 12 unique GO terms associated with 7 genes for R12A × R12C, and 9 matches to 9 unique GO terms associated with 9 genes for LA × PRC (Table S5). We did not perform additional statistical tests due to the limited sample size for each GO term.

Finally, we conducted annotation and tests for functional enrichment for the highly divergent SNPs that were non-parallel (i.e., in the tail of the empirical F_{ST} distribution only for a single population pair). We delimited this set of SNPs such that their numbers would be comparable to the 439 most parallel SNPs discussed in the main text (in particular we selected the 110 SNPs or 0.0025% with the highest estimate of F_{ST} in each population pair which gave us 440 unique SNPs, i.e., each of these SNPs had this high level of F_{ST} in only a single population pair). These are our strongest candidates for non-parallel but highly divergent SNPs. Similarly to the top parallel SNPs datasets, a randomization approach was followed to test for functional enrichment. As before, we retrieved the closest predicted gene within the same scaffold, and collected the GO terms that mapped to the Pfam domains that matched that gene. In this dataset, we found 227 matches to 78 unique GO terms associated with 58 SNPs. In particular, we found 139 matches to 41 unique molecular function terms associated with 53 SNPs, 24 matches to 10 unique cellular component terms associated with 20 SNPs, and 64 matches to 27 unique biological process terms associated with 37 SNPs (Table S6). As previously, we calculated Bonferroni-adjusted 2-tail p-values for each unique GO term by comparing the proportion of associated SNPs observed to the expected empirical null distribution. Likewise, we excluded the unique GO terms where the number of associated SNPs was ≤ 2.

Supplementary Text

Data and computer source code

Raw sequencing reads are available from the NCBI short read archive (BioProject ID:PRJNA243533). Whole genome assembly and annotations can be downloaded from http://nosil-lab.group.shef.ac.uk/resources. Additional data and computer source code have been deposited in the Dryad repository doi:10.5061/dryad.74j0, and are also available from the authors upon request.

References and Notes

1. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013). Medline doi:10.1038/nrg3483
2. R. D. H. Barrett, H. E. Hoekstra, Molecular spandrels: Tests of adaptation at the genetic level. *Nat. Rev. Genet.* **12**, 767–780 (2011). Medline doi:10.1038/nrg3015
3. M. K. Burke, How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proc. R. Soc. London Ser. B* **279**, 5029–5038 (2012). Medline doi:10.1098/rspb.2012.0799
4. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006). Medline doi:10.1126/science.1123539
5. J. R. Meyer, D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, R. E. Lenski, Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* **335**, 428–432 (2012). Medline doi:10.1126/science.1214449
6. J. B. Losos, *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles* (Univ. California Press, Berkeley, CA, 2009).
7. F. C. Jones, M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, E. S. Lander, F. Di Palma, K. Lindblad-Toh, D. M. Kingsley, The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012). Medline doi:10.1038/nature10944
8. G. L. Conte, M. E. Arnegard, C. L. Peichel, D. Schluter, The probability of genetic parallelism and convergence in natural populations. *Proc. R. Soc. London Ser. B* **279**, 5039–5047 (2012). Medline doi:10.1098/rspb.2012.2146

9. Heliconius Genome Consortium, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012). Medline

10. D. Schluter, L. M. Nagel, Parallel speciation by natural selection. *Am. Nat.* **146**, 292–301 (1995). doi:10.1086/285799

11. J. L. Feder, S. P. Egan, P. Nosil, The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012). Medline doi:10.1016/j.tig.2012.03.009

12. H. Ellegren, L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H. Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, S. Uebbing, J. B. Wolf, The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760 (2012). Medline

13. S. H. Martin, K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, C. D. Jiggins, Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013). Medline doi:10.1101/gr.159426.113

14. H. Ellegren, Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29**, 51–63 (2014). Medline doi:10.1016/j.tree.2013.09.008

15. P. Nosil, Divergent host plant adaptation and reproductive isolation between ecotypes of *Timema cristinae* walking sticks. *Am. Nat.* **169**, 151–162 (2007). Medline doi:10.1086/510634

16. P. Nosil, Z. Gompert, T. E. Farkas, A. A. Comeault, J. L. Feder, C. A. Buerkle, T. L. Parchman, Genomic consequences of multiple speciation processes in a stick insect. *Proc. R. Soc. London Ser. B* **279**, 5058–5065 (2012). Medline doi:10.1098/rspb.2012.0813

17. P. Nosil, B. J. Crespi, C. P. Sandoval, Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature* **417**, 440–443 (2002). Medline doi:10.1038/417440a

18. Z. Gompert, A. A. Comeault, T. E. Farkas, J. L. Feder, T. L. Parchman, C. A. Buerkle, P. Nosil, Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* **17**, 369–379 (2014). Medline doi:10.1111/ele.12238

19. Materials and methods are available as supplementary materials on *Science* Online.

20. M. K. N. Lawniczak, S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson, B. White, S. Redmond, L. Fulton, E. Appelbaum, J. Godfrey, C. Farmer, A. Chinwalla, S. P. Yang, P. Minx, J. Nelson, K. Kyung, B. P. Walenz, E. Garcia-Hernandez, M. Aguiar, L. D. Viswanathan, Y. H. Rogers, R. L. Strausberg, C. A. Saski, D. Lawson, F. H. Collins, F. C. Kafatos, G. K. Christophides, S. W. Clifton, E. F. Kirkness, N. J. Besansky, Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* **330**, 512–514 (2010). Medline doi:10.1126/science.1195755

21. S. Renaut, C. J. Grassa, S. Yeaman, B. T. Moyers, Z. Lai, N. C. Kane, J. E. Bowers, J. M. Burke, L. H. Rieseberg, Genomic islands of divergence are not affected by

geography of speciation in sunflowers. *Nature Communications* **4**, 1827 (2013). Medline doi:10.1038/ncomms2833

22. T. Hofer, M. Foll, L. Excoffier, Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC Genomics* **13**, 107 (2012). Medline doi:10.1186/1471-2164-13-107

23. T. van Ooik, M. J. Rantala, Local adaptation of an insect herbivore to a heavy metal contaminated environment. *Ann. Zool. Fenn.* **47**, 215–222 (2010). doi:10.5735/086.047.0306

24. J. F. V. Vincent, *Structural Biomaterials* (Princeton Univ. Press, Princeton, NJ, 1990).

25. N. T. Dittmer, M. R. Kanost, Insect multicopper oxidases: diversity, properties, and physiological roles. *Insect Biochem. Mol. Biol.* **40**, 179–188 (2010). Medline doi:10.1016/j.ibmb.2010.02.006

26. E. E. Hare, J. S. Johnston, "Genome size determination using flow cytometry of propidium iodide-stained nuclei," in *Molecular Methods for Evolutionary Genetics*, V. Orgogozo, M. V. Rockman, Eds. (Humana, New York, 2011), vol. 772, pp. 3–12.

27. R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, J. Wang, *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010). Medline doi:10.1101/gr.097261.109

28. S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, D. B. Jaffe, High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1513–1518 (2011). Medline doi:10.1073/pnas.1017351108

29. A. A. Comeault, M. Sommers, T. Schwander, C. A. Buerkle, T. E. Farkas, P. Nosil, T. L. Parchman, De novo characterization of the *Timema cristinae* transcriptome facilitates marker discovery and inference of genetic divergence. *Mol. Ecol. Resour.* **12**, 549–561 (2012). Medline doi:10.1111/j.1755-0998.2012.03121.x

30. M. Yandell, D. Ence, A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012). Medline doi:10.1038/nrg3174

31. C. Holt, M. Yandell, MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011). Medline doi:10.1186/1471-2105-12-491

32. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0, 1996-2013 (2013); www.repeatmasker.org.

33. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). Medline doi:10.1093/nar/27.2.573

34. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005). Medline doi:10.1159/000084979

35. T. J. Wheeler, J. Clements, S. R. Eddy, R. Hubley, T. A. Jones, J. Jurka, A. F. Smit, R. D. Finn, Dfam: A database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41**, D70–D82 (2013). Medline doi:10.1093/nar/gks1265

36. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004). Medline doi:10.1186/1471-2105-5-59

37. V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, M. Borodovsky, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008). Medline doi:10.1101/gr.081612.108

38. G. Parra, K. Bradnam, I. Korf, CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007). Medline doi:10.1093/bioinformatics/btm071

39. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). Medline doi:10.1186/1471-2105-10-421

40. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004). Medline doi:10.1101/gr.1865504

41. S. R. Eddy, “A new generation of homology search tools based on probabilistic inference,” in *Genome Informatics 2009*, S. Morishita *et al*., Eds. (Genome Informatics Series, Imperial College Press, London, 2009), vol. **23**, pp. 205–211.

42. G. Parra, E. Blanco, R. Guigó, GeneID in *Drosophila*. *Genome Res.* **10**, 511–515 (2000). Medline doi:10.1101/gr.10.4.511

43. M. Magrane, U. Consortium, UniProt Knowledgebase: A hub of integrated protein data. *Database* **2011**, bar009 (2011). Medline doi:10.1093/database/bar009

44. H. O. Letsch, K. Meusemann, B. Wipfler, K. Schütte, R. Beutel, B. Misof, Insect phylogenomics: results, problems and the impact of matrix composition. *Proc. Biol. Sci.* **279**, 3282–3290 (2012). Medline doi:10.1098/rspb.2012.0744

45. J. Falgueras, A. J. Lara, N. Fernández-Pozo, F. R. Cantón, G. Pérez-Trabado, M. G. Claros, SeqTrim: A high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* **11**, 38 (2010). Medline doi:10.1186/1471-2105-11-38

46. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). Medline doi:10.1038/nbt.1883

47. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005). Medline doi:10.1186/1471-2105-6-31

48. K. Eilbeck, B. Moore, C. Holt, M. Yandell, Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* **10**, 67 (2009). Medline doi:10.1186/1471-2105-10-67

49. E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, R. Lopez, InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33** (suppl. 2), W116–W120 (2005). Medline doi:10.1093/nar/gki442

50. R. Li, J. Ye, S. Li, J. Wang, Y. Han, C. Ye, J. Wang, H. Yang, J. Yu, G. K. Wong, J. Wang, ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLOS Comput. Biol.* **1**, e43 (2005). Medline doi:10.1371/journal.pcbi.0010043

51. A. L. Price, N. C. Jones, P. A. Pevzner, *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005). Medline doi:10.1093/bioinformatics/bti1018

52. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). Medline doi:10.1093/bioinformatics/btq461

53. C. Feschotte, U. Keswani, N. Ranganathan, M. L. Guibotsy, D. Levine, Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.* **1**, 205–220 (2009). Medline doi:10.1093/gbe/evp023

54. J. Jurka, W. Bao, K. K. Kojima, Families of transposable elements, population structure and the origin of species. *Biol. Direct* **6**, 44 (2011). Medline doi:10.1186/1745-6150-6-44

55. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). Medline doi:10.1093/bioinformatics/btp324

56. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). Medline doi:10.1093/bioinformatics/btp352

57. T. Jombart, S. Devillard, F. Balloux, Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010). Medline doi:10.1186/1471-2156-11-94

58. Y. D. Tan, Y. X. Fu, A novel method for estimating linkage maps. *Genetics* **173**, 2383–2390 (2006). Medline doi:10.1534/genetics.106.057638

59. A. Stamatakis, A. J. Aberer, paper presented at the 27th IEEE International on Parallel and Distributed Processing Symposium (IPDPS), Boston, MA, 20 to 24 May 2013.

60. A. Stamatakis, RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006). Medline doi:10.1093/bioinformatics/btl446

61. M. Simonsen, T. Mailund, C. N. Pedersen, in *Algorithms in Bioinformatics: 8th International Workshop, WABI 2008, Karlsruhe, Germany, September 2008 Proceedings*, K. Crandall, J. Lagergren, Eds. (vol. 5251 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2008), pp. 113–122.

62. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2—approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010). Medline doi:10.1371/journal.pone.0009490

63. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005). doi:10.1111/j.1096-0031.2005.00059.x

64. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). Medline

65. H. Shimodaira, M. Hasegawa, CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001). Medline doi:10.1093/bioinformatics/17.12.1246

66. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002). Medline doi:10.1080/10635150290069913

67. R Development Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2013); www.R-project.org.

68. S. Dray, A.-B. Dufour, The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).

69. R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992). Medline

70. G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price, Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013). Medline doi:10.1101/gr.154831.113

71. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011). Medline doi:10.1093/bioinformatics/btr509

72. L. E. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41**, 164–171 (1970). doi:10.1214/aoms/1177697196

73. D. Harte, HiddenMarkov: Hidden Markov Models (2012); http://cran.r-project.org/web/packages/HiddenMarkov/index.html.

74. P. Nosil, B. J. Crespi, Experimental evidence that predation promotes divergence in adaptive radiation. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9090–9095 (2006). Medline doi:10.1073/pnas.0601575103

75. C. P. Sandoval, The effects of relative geographical scales of gene flow and selection on morph frequencies in the walking-stick *Timema cristinae*. *Evolution* **48**, 1866–1879 (1994). doi:10.2307/2410514

76. P. Nosil, Reproductive isolation caused by visual predation on migrants between divergent environments. *Proc. R. Soc. London Ser. B* **271**, 1521–1528 (2004). Medline doi:10.1098/rspb.2004.2751

77. C. Sandoval, Persistence of a walking-stick population (Phasmatoptera: Timematodea) after a wildfire. *Southwest. Nat.* **45**, 123–127 (2000). doi:10.2307/3672452

78. O. E. Gaggiotti, M. Foll, Quantifying population structure using the F-model. *Mol. Ecol. Resour.* **10**, 821–830 (2010). Medline doi:10.1111/j.1755-0998.2010.02873.x

79. D. Falush, M. Stephens, J. K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003). Medline

80. G. Nicholson, A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson, P. Donnelly, Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 695–715 (2002). doi:10.1111/1467-9868.00357

81. M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, F. Rossi, *GNU Scientific Library Reference Manual* (Network Theory Limited, Bristol, UK, ed. 3, 2009).

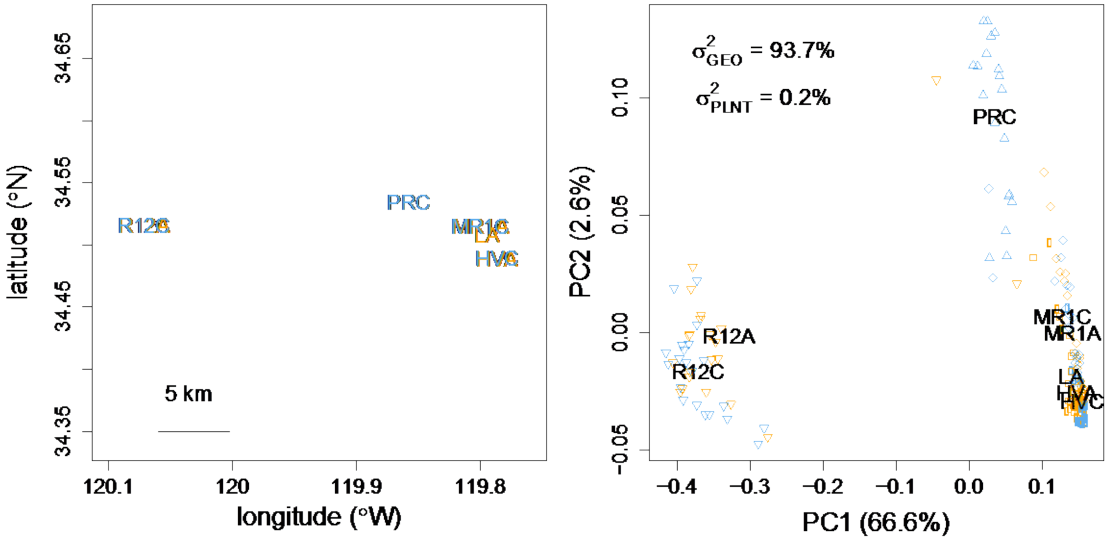

Fig. S1.

A map of the study populations (left) and the results of principal components analysis of the whole genome re-sequencing data (right). Populations group genetically strongly according to geography. The top left section of the right hand panel shows variance partitioning of PC1 by geography (GEO) versus host plant (PLNT).

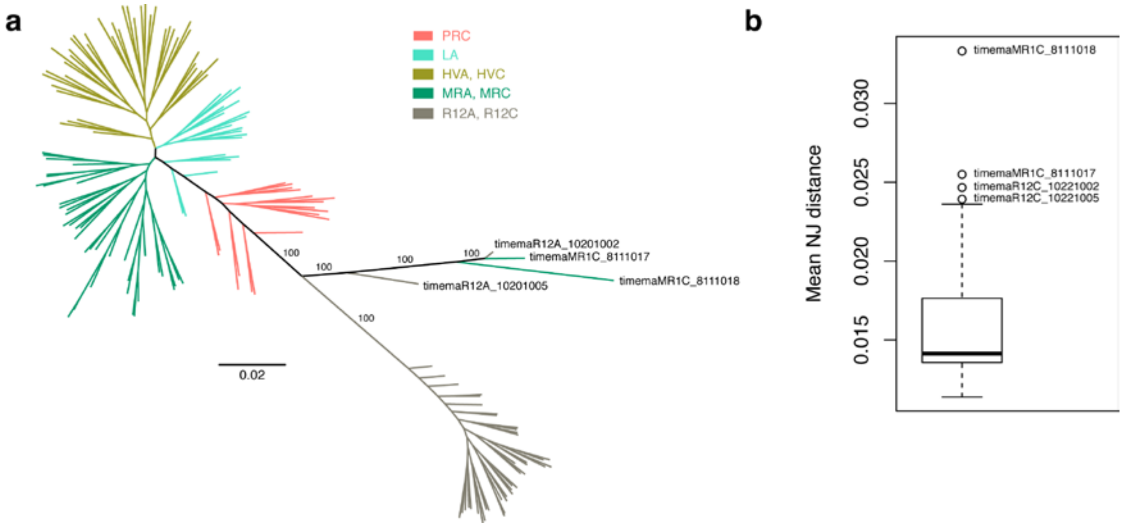

Fig. S2

Unrooted *T. cristinae* tree. a) Unrooted phylogenetic tree of 160 *T. cristinae* individuals. Note strong grouping by geography (depicted by colors of branches). Only names for the four individuals that grouped together due to a putative long-branch attraction artifact are shown. b) Boxplot of mean neighbor-joining distances between each individual and the rest.

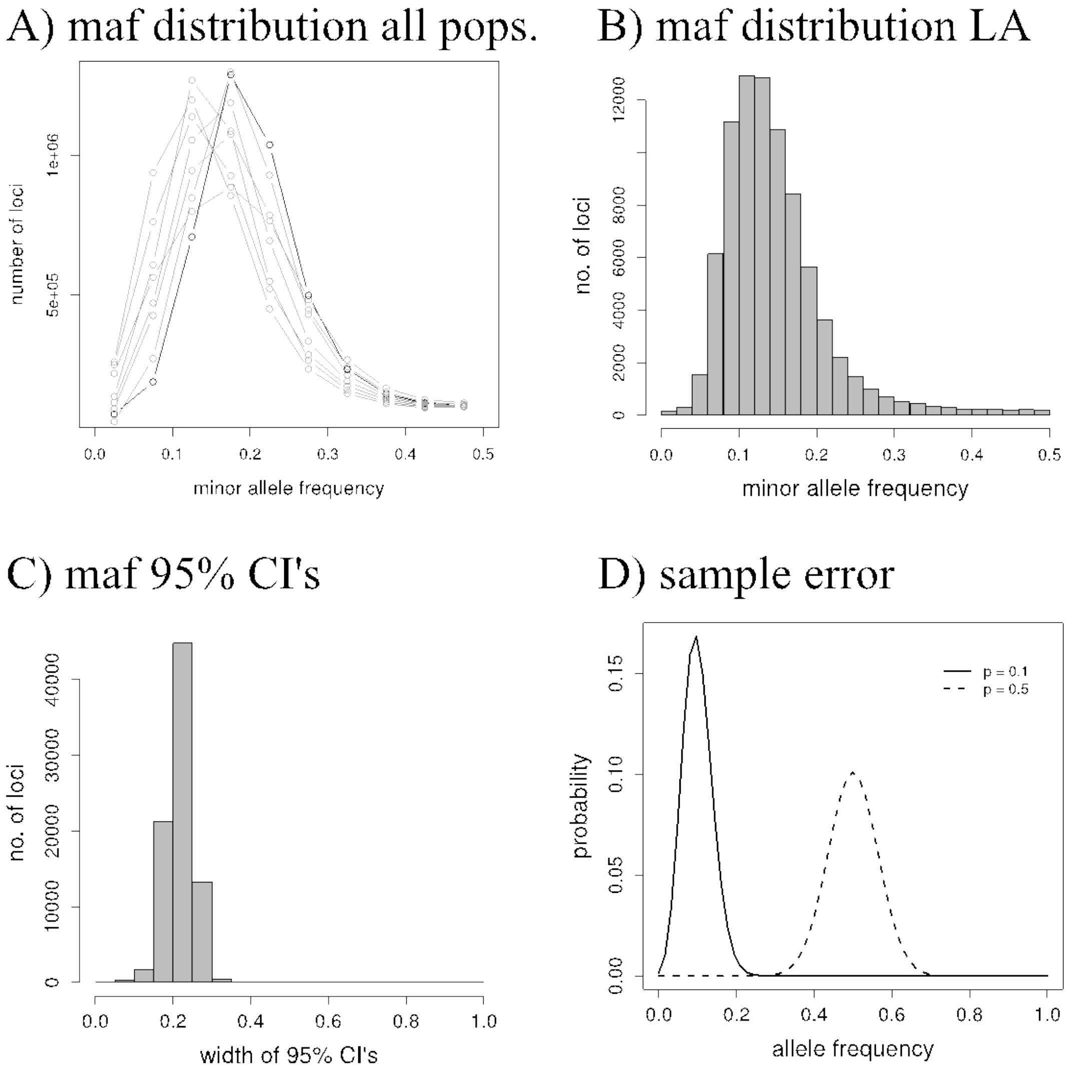

Fig. S3

Summary of genetic variation. a) The minor allele frequency (maf) distribution in each of the eight study populations (pops) examined here. Note that the individuals used in the experimental transplant were collected from population LA. b) Genetic variation in the 31 individuals used to estimate ancestral allele frequencies in the field transplant experiment. c) Width of the empirical 95% credible intervals (C.I.) for the allele frequency estimates for the SNPs examined in the experiment based on the Wright-Fisher model. d) Estimates of the expected error in allele frequency estimates based on a binomial probability distribution given the sample size (n = 31 diploid individuals) with a true allele frequency of 10% or 50%.

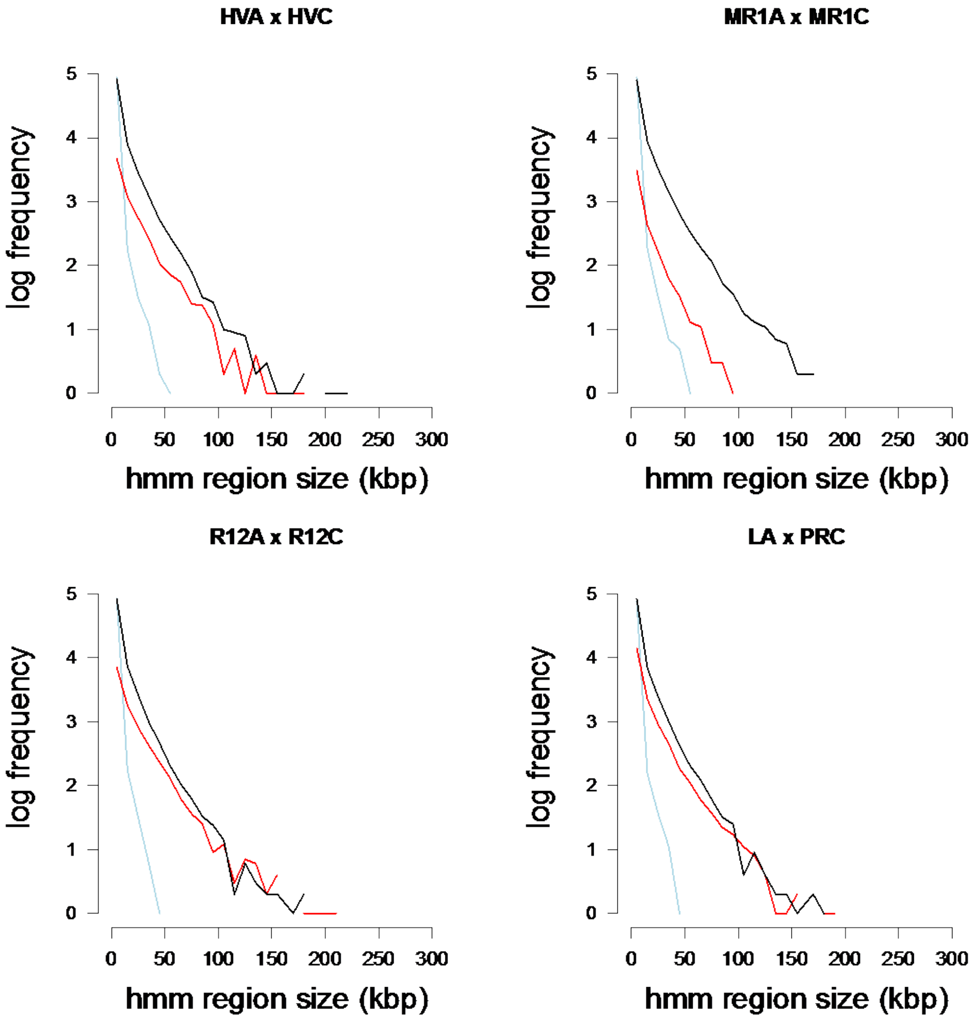

Fig. S4

Distribution of the size of HMM regions (colors are as in the main text: blue = low differentiation regions, black = moderate differentiation regions, red/orange = high differentiation regions). Absolute frequencies across the three categories are shown.

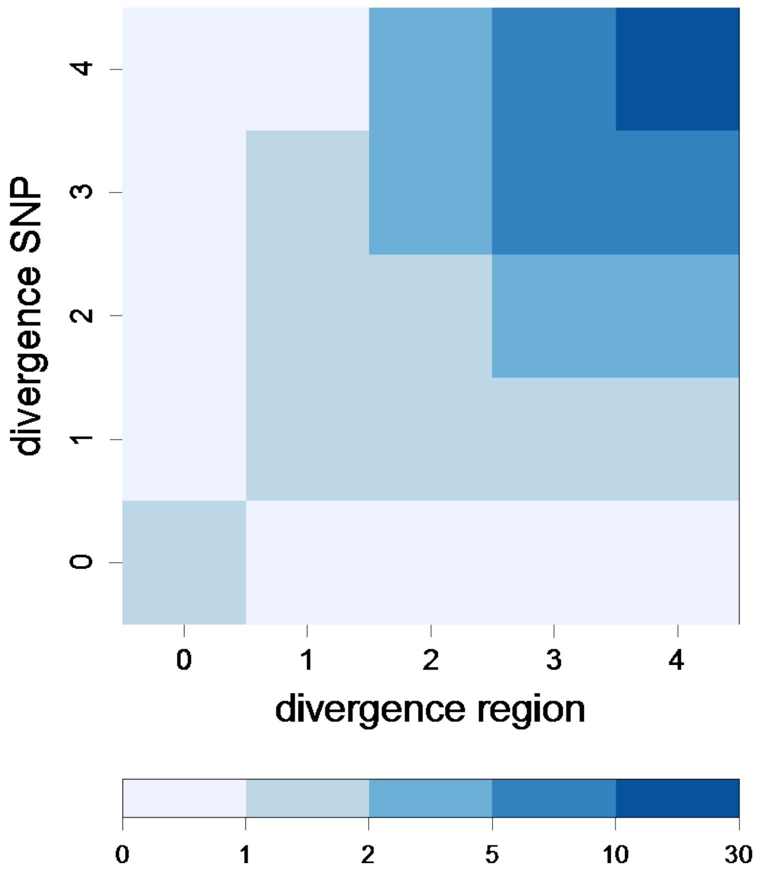

Fig. S5

Association between divergent SNPs and HMM divergence regions. The numbers on the axes (from zero to four) indicate the number of times a SNP was classified as high divergence between zero to four population pairs (y-axis) and in a HMM divergence region between zero to four population pairs (x-axis). Shading represents the ratio of observed values relative to that expected if divergent SNPs and HMM divergence regions were independent from one another. Thus, darker shading represents a stronger correspondence between divergent SNPs and HMM divergence regions. See Table S2 for raw values and significance tests.

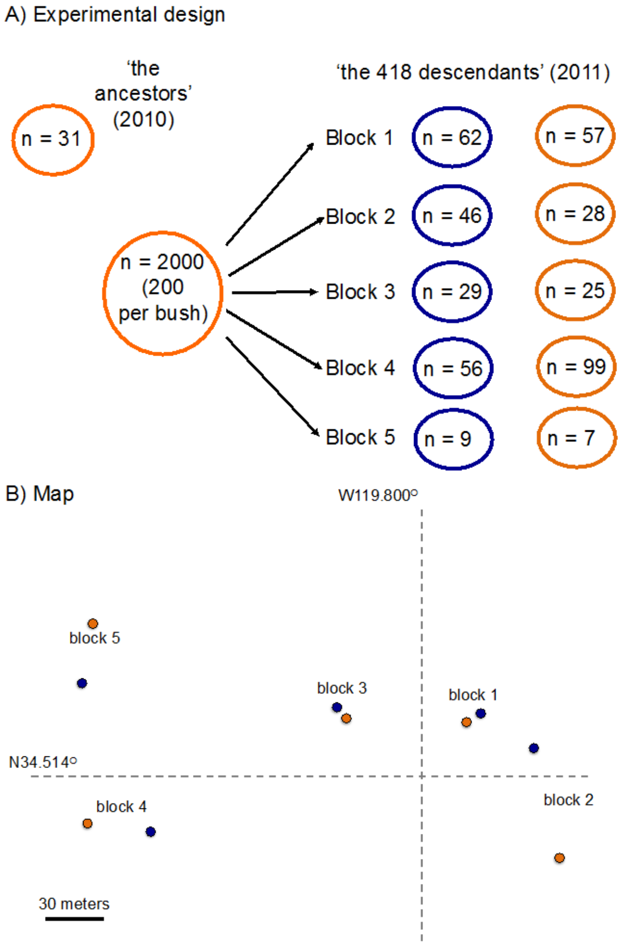

Fig. S6

Depiction of the experimental set-up. a) A total of 2000 *T. cristinae* were transplanted in 2010 to plant individuals devoid of *T. cristinae*, half onto *Ceanothus* (blue circles) and half onto *Adenostoma* (orange circles), in a paired-block design (n = 200 per individual bush). The source population was LA where *T. cristinae* feed on *Adenostoma*. A sample of 31 individuals from the same founding population as the released individuals was preserved ('the ancestors'). The descendants of the released insects were collected one generation later (2011). Numbers collected in 2011 on each experimental bush are indicated within colored circles. b) A map of the experimental site.

Table S1.

Test for parallelism of highly divergent SNPs. Quantiles refer to the empirical F_{ST} distribution. See also Figure 2 in the main text.

	Observed number of SNPs	**Expected number of SNPs**	**p-value**	**Enrichment (observed / expected)**
>90th quantile in 2 population pairs	228,347	213,423	<0.001	1.07
>90th quantile in 3 population pairs	22,366	15,809	<0.001	1.41
>90th quantile in 4 population pairs	945	439	<0.001	2.15
>90th quantile in 2 or more population pairs	251,658	229,671	<0.001	1.10
>99th quantile in 2 population pairs	4757	2582	<0.001	1.84
>99th quantile in 3 population pairs	94	17	<0.001	5.53
>99th quantile in 4 population pairs	0	0	n/a	n/a
>99th quantile in 2 or more population pairs	4851	2600	<0.001	1.87

Table S2.

Correspondence of results obtained from divergent SNPs and HMM divergence regions. The table compares the number of times a SNP was classified as high divergence (i.e., above >90th empirical quantile of the F_{ST} distribution) between zero to four population pairs (rows) and in a HMM divergence region between zero to four population pairs (columns). Shown are observed numbers, followed in parentheses by the numbers expected if SNP and HMM region results were independent from one another. Observed values significantly greater than expected at $p < 0.001$ are denoted in bold font. Note the significant results along the diagonal indicating parallel divergence SNPs tend to be in parallel HMM divergence regions more often than expected by chance. See also Figure S5.

		Number of times a SNP was in a HMM divergence region between population pairs (0-4)				
		0	1	2	3	4
Number of times a SNP was classified as high F_{ST} divergence between population pairs (0-4)	0	**1,508,297** **(1,349,886)**	461,752 (569,965)	61,859 (105,894)	4517 (10,372)	131 (438)
	1	1,082,443 (1,145,473)	**536,086** **(483,655)**	**100,128** **(89,859)**	**9163** **(8801)**	341 (372)
	2	284,357 (360,480)	**198,846** **(152,206)**	**54,004** **(28,279)**	**6334** **(2770)**	**310** **(117)**
	3	34,086 (51,748)	**30,404** **(21,850)**	**11,334** **(4059)**	**2109** **(398)**	**138** **(17)**
	4	1663 (3259)	**1964** **(1376)**	**1022** **(256)**	**243** **(25)**	**25** **(1)**

Table S3.

Parallel divergence SNPs in coding sequences of genes. Numbers of SNPs found in genic regions (ngene) and in coding sequence regions (ncds) for the most parallel SNPs under various metrics of parallelism and for various quantile cut-offs, and the same information for all SNPs and the genome as a whole (for the latter, the numbers represent bp, not SNPs).

set	**ntotal**	**ngene**	**ncds**
most parallel SNPs			
AFD Quantile = 0.00001	44	10	4
AFD Quantile = 0.000025	110	32	15
AFD Quantile = 0.00005	220	54	25
AFD Quantile = 0.0001	439	111	44
F_{ST} Quantile = 0.00001	44	17	5
F_{ST} Quantile = 0.000025	110	38	12
F_{ST} Quantile = 0.00005	220	70	21
F_{ST} Quantile = 0.0001	439	138	47
Combined Quantile = 0.00001	44	9	5
Combined Quantile = 0.000025	110	30	13
Combined Quantile = 0.00005	220	61	27

Combined Quantile = 0.0001	439	122	56
all SNPs	4,391,556	1,174,046	361,388
genome as a whole	844,841,299	190,238,410	32,233,651

Table S4.

Gene ontology (GO) terms for parallel divergence SNPs (0.001 quantile of the combined metric of parallelism). Annotations are based on the nearest gene to each parallel divergence SNP, with the requirement that the nearest gene is on the same scaffold as each focal SNP. Several GO terms can map to the same gene and therefore the sum of SNPs that match to each GO term shown here (122 for molecular function, 11 for cellular component, and 49 for biological process) is higher than the number of SNPs associated with at least one GO term (41 for molecular function, 9 for cellular component and 28 for biological process). Statistically enriched functions ($p < 0.05$ after correction for multiple comparisons) are shown in bold italics (only tested for functions represented by >2 SNPs).

Number of SNPs	GO	Function
Molecular function (41 SNPs)		
24	GO:0005515	protein binding
20	***GO:0046872***	***metal ion binding***
13	***GO:0005509***	***calcium ion binding***
10	GO:0005524	ATP binding
6	GO:0003677	DNA binding
5	GO:0016491	oxidoreductase activity
4	GO:0003824	catalytic activity
4	GO:0005525	GTP binding
3	GO:0003676	nucleic acid binding
2	GO:0003924	GTPase activity
2	GO:0003887	DNA-directed DNA polymerase activity
2	GO:0004550	nucleoside diphosphate kinase activity
2	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
2	GO:0016887	ATPase activity
2	GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
2	GO:0043169	cation binding
2	GO:0008270	zinc ion binding
2	GO:0042626	ATPase activity, coupled to transmembrane movement of substances
1	GO:0016791	phosphatase activity
1	GO:0004334	fumarylacetoacetase activity
1	GO:0004672	protein kinase activity
1	GO:0004812	aminoacyl-tRNA ligase activity
1	GO:0000166	nucleotide binding
1	GO:0001104	RNA polymerase II transcription cofactor activity
1	GO:0003964	RNA-directed DNA polymerase activity
1	GO:0046873	metal ion transmembrane transporter activity
1	GO:0016301	kinase activity
1	GO:0004221	ubiquitin thiolesterase activity
1	GO:0003777	microtubule motor activity
1	GO:0003678	DNA helicase activity
1	GO:0003723	RNA binding
1	GO:0048037	cofactor binding
1	GO:0051287	NAD binding
Cellular component (9 SNPs)		

6	GO:0016021	integral to membrane
1	GO:0016020	membrane
1	GO:0005643	nuclear pore
1	GO:0030286	dynein complex
1	GO:0005634	nucleus
1	GO:0016592	mediator complex

Biological process (28 SNPs)

7	GO:0055085	transmembrane transport
4	GO:0008152	metabolic process
4	GO:0055114	oxidation-reduction process
3	GO:0005975	carbohydrate metabolic process
2	GO:0006165	nucleoside diphosphate phosphorylation
2	GO:0007165	signal transduction
2	GO:0006520	cellular amino acid metabolic process
2	GO:0006183	GTP biosynthetic process
2	GO:0006810	transport
2	GO:0006228	UTP biosynthetic process
2	GO:0006260	DNA replication
2	GO:0006241	CTP biosynthetic process
1	GO:0006418	tRNA aminoacylation for protein translation
1	GO:0009072	aromatic amino acid family metabolic process
1	GO:0006511	ubiquitin-dependent protein catabolic process
1	GO:0007264	small GTPase mediated signal transduction
1	GO:0030001	metal ion transport
1	GO:0016973	poly(A)+ mRNA export from nucleus
1	GO:0044267	cellular protein metabolic process
1	GO:0006468	protein phosphorylation
1	GO:0005976	polysaccharide metabolic process
1	GO:0015074	DNA integration
1	GO:0006357	regulation of transcription from RNA polymerase II promoter
1	GO:0007018	microtubule-based movement
1	GO:0006281	DNA repair
1	GO:0006278	RNA-dependent DNA replication
1	GO:0000723	telomere maintenance

Table S5.

Molecular function gene ontology (GO) terms for the 32 HMM divergence regions that were significantly enriched in the SNPs showing the strongest parallel and divergent allele frequency change between hosts in the transplant experiment. Annotations are based on all the genes that overlap the 32 HMM divergence regions. Several GO terms can map to the same gene and therefore the sum of genes that match each GO term shown here (11 for HVA × HVC, 14 for MR1A × MR1C, 14 for R12A × R12C, and 9 for LA × PRC) can be higher than the number of genes associated with at least one GO term (8 for HVA × HVC, 8 for MR1A × MR1C, 7 for R12A × R12C, and 9 for LA × PRC).

		Number of genes				
GO	**Molecular function**	**Total**	**HVA × HVC**	**MR1A × MR1C**	**R12A × R12C**	**LA × PRC**
GO:0005524	ATP binding	7	2	3	2	0
GO:0016301	kinase activity	6	2	2	2	0
GO:0008234	cysteine-type peptidase activity	4	1	1	1	1
GO:0015078	hydrogen ion transmembrane transporter activity	4	1	1	1	1
GO:0003735	structural constituent of ribosome	4	1	1	1	1
GO:0005509	calcium ion binding	4	1	1	1	1
GO:0005215	transporter activity	4	1	1	1	1
GO:0046983	protein dimerization activity	3	1	1	0	1
GO:0042302	structural constituent of cuticle	2	0	1	1	1
GO:0008199	ferric iron binding	2	0	1	0	1
GO:0005525	GTP binding	2	0	0	1	1
GO:0003677	DNA binding	1	1	0	0	0
GO:0004672	protein kinase activity	1	0	1	0	0
GO:0005506	iron ion binding	1	0	0	1	0
GO:0016491	oxidoreductase activity	1	0	0	1	0
GO:0003824	catalytic activity	1	0	0	1	0

Table S6.

Gene ontology (GO) terms for divergent SNPs restricted to a single population pair (i.e., non-parallel divergence SNPs). Annotations are based on the nearest gene to each parallel divergence SNP, with the requirement that the nearest gene is on the same scaffold as each focal SNP. Several GO terms can map to the same gene and therefore the sum of SNPs that match to each GO term shown here (139 for molecular function, 24 for cellular component, and 64 for biological process) is higher than the number of SNPs associated with at least one GO term (53 for molecular function, 20 for cellular component and 37 for biological process). Statistically enriched functions ($p < 0.05$ after correction for multiple comparisons) are shown in bold italics (only tested for functions represented by >2 SNPs).

Number of SNPs	**GO**	**Function**
Molecular function (53 SNPs)		
38	GO:0005515	protein binding
12	***GO:0008270***	***zinc ion binding***
8	GO:0005524	ATP binding
6	***GO:0004812***	***aminoacyl-tRNA ligase activity***
6	***GO:0043565***	***sequence-specific DNA binding***
6	GO:0003700	sequence-specific DNA binding transcription factor activity
5	GO:0003964	RNA-directed DNA polymerase activity
5	GO:0003723	RNA binding
4	GO:0003677	DNA binding
4	GO:0000166	nucleotide binding
4	GO:0005509	calcium ion binding
4	GO:0016787	hydrolase activity
3	GO:0003824	catalytic activity
2	GO:0008237	metallopeptidase activity
2	GO:0003676	nucleic acid binding
2	GO:0004842	ubiquitin-protein ligase activity
2	GO:0005507	copper ion binding
2	GO:0003916	DNA topoisomerase activity
2	GO:0046872	metal ion binding
1	GO:0031683	G-protein beta/gamma-subunit complex binding
1	GO:0004672	protein kinase activity
1	GO:0004871	signal transducer activity
1	GO:0003735	structural constituent of ribosome
1	GO:0004827	proline-tRNA ligase activity
1	GO:0016746	transferase activity, transferring acyl groups
1	GO:0003924	GTPase activity
1	GO:0003887	DNA-directed DNA polymerase activity

1	GO:0004970	ionotropic glutamate receptor activity
1	GO:0042302	structural constituent of cuticle
1	GO:0008408	3'-5' exonuclease activity
1	GO:0008061	chitin binding
1	GO:0015116	sulfate transmembrane transporter activity
1	GO:0005234	extracellular-glutamate-gated ion channel activity
1	GO:0016491	oxidoreductase activity
1	GO:0005086	ARF guanyl-nucleotide exchange factor activity
1	GO:0046983	protein dimerization activity
1	GO:0046873	metal ion transmembrane transporter activity
1	GO:0019001	guanyl nucleotide binding
1	GO:0003678	DNA helicase activity
1	GO:0005215	transporter activity
1	GO:0004930	G-protein coupled receptor activity

Cellular component (20 SNPs)

8	GO:0016020	membrane
7	***GO:0005634***	***nucleus***
2	GO:0016021	integral component of membrane
1	GO:0005737	cytoplasm
1	GO:0005840	ribosome
1	GO:0000922	spindle pole
1	GO:0005576	extracellular region
1	GO:0005815	microtubule organizing center
1	GO:0005694	chromosome
1	GO:0032040	small-subunit processome

Biological process (37 SNPs)

6	***GO:0006418***	***tRNA aminoacylation for protein translation***
6	GO:0006355	regulation of transcription, DNA-templated
5	GO:0008152	metabolic process
5	GO:0006278	RNA-dependent DNA replication
4	***GO:0046080***	***dUTP metabolic process***
4	GO:0007156	homophilic cell adhesion
4	GO:0015074	DNA integration
2	GO:0006265	DNA topological change
2	GO:0006457	protein folding
2	GO:0007186	G-protein coupled receptor signaling pathway
1	GO:0006412	translation
1	GO:0006030	chitin metabolic process
1	GO:0055114	oxidation-reduction process
1	GO:0006810	transport
1	GO:0000226	microtubule cytoskeleton organization

1	GO:0006260	DNA replication
1	GO:0030001	metal ion transport
1	GO:0008272	sulfate transport
1	GO:0006468	protein phosphorylation
1	GO:0006433	prolyl-tRNA aminoacylation
1	GO:0055085	transmembrane transport
1	GO:0006281	DNA repair
1	GO:0032012	regulation of ARF protein signal transduction
1	GO:0000723	telomere maintenance