Supplementary Information—Supple et al. 2013

Table of Contents

S1. Annotation of the *H. erato* Red Color-Pattern (*D*) Interval

S1.1 Annotation Methods

We annotated genes across the red "*D*" color pattern interval of *H. erato*. To provide supporting evidence for gene models, we sequenced and aligned short-read transcriptome data from several races and stages of *H. erato* to the available partial genomic reference sequence. Additional supporting evidence was provided by aligning available EST and protein databases. We manually curated the predicted genes and compared them to predicted genes in the *H. melpomene* genome v1.1 (Heliconius Genome Consortium 2012).

Reference sequence

We examined 2.2 Mb of the approximately 400 Mb *H. erato* genome. A 1 Mb genomic region involved in red wing color pattern (the *D* interval) was sequenced from an *H. e. petiverana* BAC library (Counterman et al. 2010). In addition, we sequenced approximately 1.2 Mb from BAC clones unlinked to red color pattern. We compiled all sequences into a single *H. erato* "reference" genome (Table S1). We masked repetitive elements in this reference using RepeatMasker v3-2-9 (Smit et al. 2010) and a *Heliconius* repetitive elements database (Heliconius Genome Consortium 2012).

Table S1: *H. erato* reference sequence contigs

NCBI accession	size (bp)	description
KC469892	144495	unlinked to red color pattern
KC469893	743824	unlinked to red color pattern
KC469894	948009	red color-pattern (*D*) interval
KC469895	100927	unlinked to red color pattern
AC208805	70206	unlinked to red color pattern
AC208806	134000	unlinked to red color pattern

Transcriptome assembly

We generated a partial reference transcriptome for *H. erato* wing tissue using reference based assembly of Illumina RNA-seq short-read data from hind wing cDNA from 18 individuals, representing two divergent color pattern races (*H. e. favorinus* and *H. e. emma*) that meet in the Peruvian hybrid zone. Each race was sampled in three biological replicates across three developmental stages (5th instar, day 1 pupae, day 3 pupae). These stages are most relevant to phenotypic differentiation, as they precede the physical manifestation of the color phenotype, which occurs around 5 days after pupation, and include the stage of initial differential expression in *optix*, which occurs at day 3 (Reed et al. 2011).

For each individual, we obtained cDNA from whole hindwing tissues and prepared libraries for sequencing using a slightly-modified Illumina protocol (Illumina 2008, outlined in Supplementary Protocols). Briefly, this involved RNA extraction and isolation of mRNA using poly-A tail binding. Transcripts were chemically fragmented and converted to cDNA using random primers. Adapters were ligated and the resulting fragments were size selected from 100-250 bp using gel extraction, and amplified using a 15-cycle PCR.

Each sample was run in a single Illumina lane and paired end sequenced at either 36, 66, or 75 bp lengths on an Illumina GAIIx at either the UNC-CH Genome Analysis Facility or NCSU

Genome Sequence Laboratories. The number of reads ranged from 19.2 to 55.9 million per sample. Sample quality was assessed using FastQC v0.8.0 (Andrews 2011) and a few low quality samples were rerun. To standardize read length and remove adapter contamination due to read-through of short fragments, all reads were trimmed to 36 bp using FASTX-Toolkit's fastx_trimmer v0.0.13 (Gordon 2010). We obtained empirical estimates of the distribution of sequenced fragment lengths for each sample by aligning the reads to the unmasked *H. erato* reference sequence using BWA v0.5.9-r16 (Li and Durbin 2009) with default parameters.

We generated transcripts with a reference-based assembly method using the Bowtie/TopHat/Cufflinks pipeline. Each sample was aligned, using the empirically estimated fragment lengths, to the masked reference sequence using the closure search option in TopHat v1.2.0 (Trapnell et al. 2009) and utilizing Bowtie v0.12.7.0 (Langmead et al. 2009). We used stringent mapping parameters to minimize false alignments (see Table S2). No more than a single mismatch was allowed per 36 bp read, each aligned read could only map to a single location, and each splice junction had to be supported by at least one read with at least 12 bases on each side with no mismatches. Intron and exon size parameters were determined by examining the distribution of sizes in *Bombyx mori* (Duan et al. 2010) and *Drosophila melanogaster* (McQuilton et al. 2012). We used a first round of TopHat alignments to create a library of potential splice junctions across all samples. This splice junction library was used when each sample was realigned using TopHat with all other parameters unchanged. To generate transcripts, alignments for all samples were first merged using SamTools v0.1.9.0 (Li et al. 2009) and then analyzed with Cufflinks v1.0.1 (Trapnell et al. 2010), using default parameters, except for decreasing the minimum and maximum exon lengths.

Table S2: TopHat and Cufflink parameters

TopHat parameter	**value**	**description**
max-multihits	1	number of alignments to allow per read
segment-mismatches	1	number of mismatches allowed per segment
min-intron-length	20	minimum intron length
max-intron-length	4000	maximum intron length
min-anchor	12	minimum number of aligned bases on each side to report a splice junction
allow-indels	true	allows indels
no-coverage-search	true	disables coverage search
closure-search	true	enables closure search
min-closure-exon	3	minimum exon length for closure search
min-closure-intron	20	minimum intron length for closure search
min-segment-intron	20	minimum intron length for split segment
max-segment-intron	20000	maximum intron length for split segment
Cufflink parameter	**value**	**description**
min-intron-length	20	minimum intron length
max-intron-length	20000	maximum intron length

Automated gene annotation

We produced automated gene annotations for the *H. erato* partial genomic reference sequence using the MAKER pipeline v2.09 (Holt and Yandell 2011) with modified parameters (Table S3).

This analysis begins with masking repetitive elements in the reference sequence using RepeatMasker v3-2-9 (Smit et al. 2010) with the *Heliconius* repetitive elements database (Heliconius Genome Consortium 2012). MAKER next aligned peptide sequences from the Uniref90 (Suzek et al. 2007) and *Bombyx mori* (Duan et al. 2010) protein databases using NCBI BLASTX v2.2.24 (Altschul et al. 1997) and polished these alignments with Exonerate v2.2.0 (Slater and Birney 2005) to ensure that multiple hits within a single protein are ordered properly and utilize consensus splice sites. MAKER then aligned *H. erato* ESTs from a previous *de novo* assembly built from Sanger and 454 sequences from wing tissues of several H. erato races and H. himera (Papanicolaou et al. 2009). These ESTs were aligned using NCBI BLASTN v2.2.24 (Altschul et al. 1997) and further polished with Exonerate v2.2.0 (Slater and Birney 2005). We included the set of aligned RNA-seq transcripts as additional EST evidence in the MAKER pipeline. MAKER next generated *ab initio* gene predictions for both the masked and unmasked reference using Augustus v2.5.5 (Stanke et al. 2006) trained for *H. melpomene* (Heliconius Genome Consortium 2012) and SNAP v2010-07-28 (Korf 2004) trained for *Bombyx mori*. Finally, MAKER determined which *ab initio* gene models had enough supporting evidence from aligned peptide sequences, ESTs, and RNA-seq to be promoted to predicted genes. We required promoted models to produce a protein with at least 30 amino acids and have an annotation edit distance (AED) no greater than 0.5, which is a measure of the difference between the model and the supporting evidence (Holt and Yandell 2011). In the event of overlapping models, only the model with the lowest AED was promoted.

Table S3: MAKER behavior options

MAKER parameter	value	description
pred_flank	500	extent of surrounding evidence to pass to gene predictors
AED_threshold	0.5	maximum annotation edit distance
min_protein	30	minimum number of amino acids in a predicted protein
alt_splice	1 [yes]	take additional steps to find alternative splicing?
always_complete	0 [no]	force start and stop codons for every gene?
keep_preds	0 [no]	include unsupported gene predictions to final gene set?
split_hit	20000	expected max intron size for alignments
single_exon	1 [yes]	include single exon EST evidence?
single_length	250	minimum length of single exon ESTs

Manual curation and functional annotation of predicted genes

We manually curated all predicted genes in the *D* interval using Apollo and following the BeeBase protocols, section IV (Munoz-Torres et al. 2011). Curation involved manually examining each predicted gene to see how well it matched the supporting evidence, adding or removing exons based on supporting evidence and shifted exon boundaries to match RNA-seq models. We blasted the resulting peptide sequences against NCBI's non-redundant (nr) protein database (http://www.ncbi.nlm.nih.gov/). Using ClustalW2 (Larkin et al. 2007), we examined alignments between top *Heliconius* hits and non-*Heliconius* hits, focusing on insect proteins from NCBI's Reference Sequence (RefSeq) collection (Pruitt et al. 2007), which is reviewed and curated. We examined alignments for major gaps or differences and attempted to modify the predicted gene to better match the top blast hits.

We assigned gene descriptions based on the top BLAST hit of the curated proteins to the SwissProt protein database (Boeckmann et al. 2003) with an e-value of at least 0.001 and Blast2Go functional annotation (Conesa et al. 2005) of the curated coding sequences blasted against NCBI's non-redundant (nr) protein database (http://www.ncbi.nlm.nih.gov/). We assigned putative functions to genes based on gene ontology terms from the Blast2Go analysis and known functions of domains identified with InterProScan domain recognition analysis (Hunter et al. 2012).

Transcription factor binding site prediction

In silico transcription factor binding site prediction was performed with the Transcription Element Search System (TESS) (Schug and Overton 1997) using the default settings, searching against known *Drosophila* binding sites from the TRANSFAC and JASPAR databases. Custom scripts were used to divide the genomic region downstream of *optix* into non-overlapping 2 kb fasta sequences for upload to TESS. Additional custom post-processing scripts were used to reassemble TESS output files for the full locus and to parse out transcription factor binding site predictions with p-values less than 0.05. The remaining predictions were uploaded to a private UCSC genome browser track for visual inspection.

S1.2 Annotation Results & Discussion

Using the MAKER pipeline, we annotated 30 protein coding genes in the *D* interval based on *ab initio* models with supporting evidence from homology to known proteins and regions of active transcription identified from ESTs and RNA-Seq data (GenBank accession KC469894). The annotated genes cover a wide variety of functions (Table S4). The distribution of genes across the *D* interval showed a 250 kb "gene desert", which contains only a single gene, *optix* (Figure 2), which has been shown to be involved in the red phenotype (Reed et al. 2011).

We used the RNA-seq alignments to identify potential non-coding transcriptional activity in the gene desert. There is transcriptional activity immediately 3' of *optix*. There are a few additional regions throughout the gene desert where RNA-seq reads aligned, but visual inspection revealed these to be artifacts due to repetitive sequences.

TESS transcription factor binding site prediction within the region of peak divergence revealed over 12,000 putative binding sites, 6,927 of which had a p-value less than 0.05. Filtering of these results for potential transcription factors associated with *optix* was unsuccessful due to a lack of known candidate binding sites. The modMine, through the *Drosophila* modEncode Project (Celniker et al. 2009), identifies *eyeless* (*ey*) as the only gene known to bind and regulate *optix*, and identifies a 1435 bp regulatory region that putatively contains the *ey* binding region. A BLAST search did not show a region of significant sequence similarity to this candidate regulatory region in our *H. erato* reference sequences. It is important to note that there is no known evidence of *optix* expression in *Drosophila* wings and it is unknown if similar genes bind and regulate *optix* in developing eyes and wings. The limited data of gene interactions with *optix* hinders our ability to identify which of the thousands of putative transcription factor binding sites across the 65 kb region may be involved in regulating *optix* expression during wing pattern development.

Table S4: Annotated coding genes, their putative functions, and *H. melpomene* orthologs

gene	description	putative function	translation start	translation stop	*H. melpomene* ortholog
HERA000001	hypothetical protein	unknown	7853	14570	HMEL003289
HERA000002	sideroflexin-2-like	cation transmembrane transporter activity	30936	19952	HMEL003292
HERA000003	PWWP domain-containing	transcription factor regulating a developmental process	34342	37275	HMEL003293
HERA000004	haspin-like	protein phosphorylation/serine threonine-protein kinase activity	54012	63492	HMEL003294
HERA000005	hypothetical protein	unknown	66376	69249	HMEL003296
HERA000006	max dimerization-like	regulation of transcription	360347	84307	HMEL001000
HERA000007	DARL anticodon-binding domain-containing	tRNA ligase activity	365880	361488	HMEL001004
HERA000008	DnaJ domain-containing	heat shock protein binding	366972	367988	HMEL001006
HERA000009	blood vessel epicardial substance-like	cell motility and cell adhesion	449332	374194	HMEL001009
HERA000010	ashwin-like	involved in embryonic morphogenesis	450061	451140	HMEL001022
HERA000011	phosphodiesterase 10a-like	involved in signal transduction	454498	464434	HMEL001021
HERA000012	sorting nexin 12-like	phosphatidylinositol binding	470177	466059	HMEL001020
HERA000013	step ii splicing factor slu7-like	Pre-mRNA splicing factor	477368	471255	HMEL001019
HERA000014	kinesin-like	microtubule motor activity	488674	480075	HMEL001018
HERA000015	G protein-coupled receptor-like	transmembrane signaling receptor activity	489421	500099	HMEL001017
HERA000016	epoxide hydrolase 4-like	hydrolase and catalytic activity	511162	505139	HMEL001014

Table S4 (cont.)

gene	description	putative function	translation start	translation stop	*H. melpomene* ortholog
HERA000017	six sine homebox transcription factor (*optix*)	regulation of transcription	680834	680031	HMEL001028
HERA000018	integrator complex subunit 7-like	subunit of the integrator complex which mediates snRNA processing	723974	754207	HMEL001044
HERA000019	leucine repeat-rich	protein binding	754592	756603	HMEL001043
HERA000020	leucine repeat-rich	protein binding	757308	760265	HMEL001042
HERA000021	strabismus/van gogh-like	involved in development	770295	764754	HMEL001039
HERA000022	monocarboxylate transporter-like	transport across membranes	771067	774473	HMEL001038
HERA000023	SCY1-like protein 2-like	protein phosphorylation/serine threonine-protein kinase activity	783021	777647	HMEL001037
HERA000024	TM2 domain-containing protein CG11103-like	unknown	783509	784039	HMEL001036
HERA000025	40S ribosomal protein S13-like	structural constituent of ribosome	786109	785302	HMEL001035
HERA000026	nadh:ubiquinone dehydrogenase-like	NADH dehydrogenase (ubiquinone) activity	786716	787140	HMEL001034
HERA000027	trafficking protein particle complex subunit 5-like	involved in vesicular transport from endoplasmic reticulum to Golgi	794553	792627	HMEL001033
HERA000028	ras-related protein rab-39b-like	involved in small GTPase mediated signal transduction	801187	798560	HMEL001031
HERA000029	THAP domain-containing	nucleic acid binding	808019	810386	HMEL001029
HERA000030	hypothetical protein	unknown	868108	868452	HMEL002053

S2. Synteny and Conservation between Co-Mimics

S2.1 Synteny and Conservation Methods

We examined gene synteny across the *D* interval between *H. erato* and *H. melpomene* genomic reference sequences. We compared the thirty curated peptide sequences from the *H. erato D* interval to the *H. melpomene* v1.0 gene set peptide sequences (Heliconius Genome Consortium 2012) using Inparanoid v4.0 (Ostlund et al. 2010) to identify one-to-one orthologs. Only matches with bootstrap support of >95% and a score of >50 were retained for analysis. We examined gene rearrangements using OrthoCluster release 2 (Vergara and Chen 2009) in rs mode. We estimated the expected number of rearrangements per Mb between *H. erato* and *H. melpomene* using a divergence time of 13.5-26.1 million years (Pohl et al. 2009) and a rearrangement rate of 0.04-0.29, which are the minimum and maximum estimates from comparisons of *H. melpomene*, *Danaus plexipus* (monarch) and *Bombyx mori* (silkworm) genome assemblies (Heliconius Genome Consortium 2012).

To determine if a genomic inversion might be present in the regions of high divergence (see below), we examined a 200 kb region with the greatest divergence between races in the *H. erato D* interval and the *H. melpomene B/D* interval. We used BreakDancer v1.2.6 (Chen et al. 2009), with default parameters, to identify regions of the reference sequence that showed paired end alignments (see below) with incorrect orientations and unexpected distances between pairs.

To examine the level of sequence conservation between *H. erato* and *H. melpomene* across the *D* interval, we used mVista LAGAN (Brudno et al. 2003) to globally align the *H. erato D* interval sequence and the *H. melpomene* scaffolds containing the orthologous genes identified by the Inparanoid analysis. We examined sequence conservation in 500 bp windows across the interval, identifying regions of greater than 90% similarity.

S2.2 Synteny and Conservation Results

Each gene in the *H. erato D* interval had an *H. melpomene* ortholog (Table S4). These orthologs identify two *H. melpomene* scaffolds (HE671887 and HE670865) orthologous to the *H. erato D* interval. The HE670865 scaffold was previously identified as the *H. melpomene B/D* interval, which is responsible for the red color phenotypes, and all genes on this scaffold had been manually curated (Heliconius Genome Consortium 2012). HE671887 was not previously identified as being adjacent to the *B/D* interval and gene HMEL003292 required manual curation. We manually curated the gene based on *H. melpomene* evidence (Heliconius Genome Consortium 2012). For the 30 *H. erato D* interval genes, gene order and orientation were completely conserved relative to *H. melpomene* (Figure S1) and all protein coding *H. melpomene* genes in the homologous regions were present in the *H. erato* annotations. Additionally, the BreakDancer analysis of read pair orientation did not highlight any inversions that could be driving elevated divergence between divergent races. Despite an expected 0.5–8.0 rearrangements/Mb between *H. erato* and *H. melpomene*, we did not detect any gene rearrangements across nearly 1 Mb of sequence across the *D* interval.

We aligned the *H. erato D* interval to the two orthologous *H. melpomene* scaffolds and examined sequence conservation across the alignment. There were 182 highly conserved regions (>90% sequence similarity in a 500 bp window), covering a total of 63 kb of sequence.

Most of the highly conserved regions (82%) fell in coding exons, covering 38 kb of sequence. Of the conserved regions not located in exons, the gene desert near *optix* contained a higher proportion—8% of the gene desert was highly conserved, while only 3% of the rest of the non-exon sequence for the entire interval was highly conserved. Several of these highly conserved regions in the gene desert contain SNPs that show perfect associations with color pattern phenotype.

S3. Sampling and Genotyping Across Replicate Hybrid Zones

S3.1 Sampling and Genotyping Methods

To determine where different red phenotypes diverge genetically, and therefore, where the genetic control of the red phenotype is most likely located, we examined genomic sequence data for multiple individuals from eight different races of *H. erato* and four races of *H. melpomene*, representing two major red phenotypes. These samples were from multiple hybrid zones, where whole genome sequencing of individuals from regions of admixture between divergent color pattern races allows fine dissection of genomic regions driving phenotypic divergence. In hybrid zones between divergent red color pattern races of *Heliconius*, the free exchange of genes will homogenize the genomes, while strong selection on the red color pattern phenotype will create peaks of genetic divergence around the genomic targets of selection.

Sampling

We collected 45 individual *H. erato* butterflies from hybrid zones in Peru, French Guiana, Ecuador, and Panama (Figure 1). Adult individuals were preserved for DNA extraction or transported live to insectaries in Gamboa, Panama to establish phenotypically pure stocks. For each of the four hybrid zones, we collected phenotypically pure samples from admixed populations where the ranges of two color pattern races overlap. In these regions of admixture, gene flow homogenizes the genomes of the two races, while strong selection on color pattern phenotype drives divergence at genomic regions responsible for color pattern phenotypes. For dissecting red color pattern variation, the hybrid zones in Peru, French Guiana, and Ecuador are considered replicate hybrid zones since each involves hybridization between rayed and postman races. The Panamanian hybrid zone serves as a control in that both races are postman phenotypes, showing variation only in the yellow phenotypic elements, which are under independent genetic control from the red elements (Mallet 1986). For each of the eight color pattern races, we collected three to eight phenotypically pure individuals.

Additionally, we collected six *H. melpomene* individuals near a hybrid zone in eastern Colombia, three samples representing each of the two major red phenotypes—the postman (*H. m. melpomene*) and the rayed (*H. m. malleti*). We assessed history across a second *H. melpomene* hybrid zone in Peru—including postman (*H. m. amaryllis*) and rayed (*H. m. aglaope*) phenotypes—using published genome resequencing data (Nadeau et al. 2012).

Sequencing and genotyping

For each sample, we extracted genomic DNA from a partial thorax or whole pupae. We prepared whole genome Illumina libraries (outlined in Supplementary Protocols). Briefly this involved shearing the DNA with a Covaris machine, followed by bead purification, and then

standard Illumina library preparation. We assessed library quality using a fluorimeter and qPCR. Whole genomes of each individual were sequenced on either an Illumina GAIIx or HiSeq at Baylor College of Medicine, producing 100 bp paired end reads. We examined sequence quality for each pair of each sample separately using FastQC v0.8.0 (Andrews 2011) and hard trimmed all reads in a set using FASTX-Toolkit's fastx_trimmer v0.0.13 (Gordon 2010) where the 25th percentile base quality score dropped below 20.

We aligned the sequencing reads to our unmasked *H. erato* reference genome using BWA v0.5.9-r16 (Li and Durbin 2009) with relaxed mapping parameters (Table S5). We assessed the quality and coverage of alignments using FlagStat and DepthOfCoverage from GATK v1.2-4 (McKenna et al. 2010, DePristo et al. 2011). We used Picard v1.53 (Broad Institute 2009) and GATK to refine the alignments by marking duplicate reads using Picard's MarkDuplicates and realigning around potential indels using GATK's RealignerTargetCreator and IndelRealigner.

We called multi-sample genotypes across samples for each race using GATK's UnifiedGenotyper with default parameters, except heterozygosity set to 0.025, and filtered genotype calls for quality using GATK's VariantFiltration, applying both site and individual sample filters (Table S5) to remove low confidence genotypes. If a site did not pass the site filtering criteria, we assigned all individuals of that race a genotype of N/N. If an individual's genotype did not pass the individual sample filtering criteria, we assigned that individual a genotype of N/N. Hypercoverage regions are indicative of repetitive elements, so based on the distribution of coverage per site for each individual, we empirically choose a hypercoverage threshold of 100x per sample.

We used the same pipeline and parameters for the *H. melpomene* Colombia data, aligning to the *H. melpomene* genome v1.1 (Heliconius Genome Consortium 2012). Additionally, we obtained unfiltered genotype calls for four individuals of each race from the *H. melpomene* hybrid zones in Peru (Nadeau et al. 2012). We filtered the genotypes using the same criteria above, with the exception of a hypercoverage cutoff of 150 due to the higher overall coverage of these samples.

Genotyping samples by aligning short sequence reads to a reference genome has inherent errors associated with it that result in incorrect genotypes. We introduced an additional source of error when we aligned whole genome reads to just a small portion of the genome. We estimated this additional error rate by aligning a single *H. timareta* sample to two *H. melpomene* reference sequences—the whole genome (v1.1) and a 2 Mb partial genome comprised of the color pattern regions. An *H. timareta* sample was used in the analysis because the amount of genetic diversity in *H. melpomene* is reduced relative to *H. erato* (Flanagan et al. 2004), while the amount of genetic diversity between *H. timareta* and *H. melpomene* is similar to that within *H. erato* (see Figure 3 in Beltrán et al. 2007). We aligned reads to the reference sequences using BWA and called genotypes with the GATK pipeline (Heliconius Genome Consortium 2012). We assumed that likely erroneous genotypes were ones that disagreed between the two methods or that were called for the reduced reference, but not the whole genome reference. We estimated the error rate as the number of likely erroneous genotypes divided by the total number of genotypes called from the partial genome alignment.

Table S5: Genotype calling parameters

BWA parameter	**value**	**description**
l	35	seed length
k	2	maximum edit in seed
n	8	maximum edits per 100 bp
o	2	maximum number of gap opens
e	3	maximum number of gap extensions
GATK parameter	**value**	**description**
heterozygosity	0.025	estimated heterozygosity
GATK filter	**value to filter out**	**description**
stand_call_conf	<30	standard minimum confidence threshold for the position, which equates to a probability of a misidentified segregating SNP of less than 0.001
DP	>100 * number of samples	hypercoverage per race
genotype GQ	<30	genotype quality for the sample, which equates to a probability of greater than 0.001 that the genotype called is incorrect
genotype DP	<10	low coverage per sample
genotype DP	>100	hypercoverage per sample
QD	<5.0	quality by depth
FS	>200	strand bias
HRun	>5	homopolymer run

S3.2 Sampling and Genotyping Results

The alignments of our *H. erato* Illumina reads to the partial genomic reference produced, on average, 75% properly paired reads—both pairs mapped in the correct orientation to each other and within the expected distance distribution. For each individual, on average, we called genotypes at 50% of the positions in our intervals overall and 56% of positions across the *D* interval (Table S6). Alignment of *H. melpomene* reads to the full *H. melpomene* reference genome produced, on average, 93% properly paired reads and 74% of genotypes called across the *B/D* interval per sample (Table S7).

Genotyping samples by aligning short sequence reads to a reference genome has inherent errors from a number of sources, including the sequencing error, alignment errors, and genotyping calling errors. These sources of error have been discussed elsewhere (Pool et al. 2010) and are affected by multiple factors, including depth of coverage. To determine the impact on error rate of mapping whole genome sequence data to only a partial genomic reference, we compared genotype calls of alignments of a single *H. timareta* individual to two different reference genomes—i) the entire *H. melpomene* reference genome (approximately 269 Mb) and ii) a 2 Mb portion of the *H. melpomene* reference genome (Table S7). This analysis suggests an additional 2.5% genotyping error rate is introduced when aligning whole genome reads to a partial genomic reference.

Table S6: Samples and sequencing data for *H. erato*

hybrid zone	race (phenotype)	sample	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	positions genotyped (%)		SNPs* per genotyped position (%)	
								all reference	*D* interval	all reference	*D* interval
Peru	favorinus (postman)	GS012	06°27'41''S 76°20'31''W	69055119	8.5	75.5	20	45.7	52.2	4.6	4.5
		NCS0471	06°28'27''S 76°00'37''W	52625225	8.1	75.2	17	43.7	50.4	4.4	4.2
		NCS0473	06°28'27''S 76°00'37''W	49703856	8.3	75.3	17	42.4	48.7	4.4	4.3
		NCS0476	06°28'27''S 76°00'37''W	59869367	7.5	75.0	21	48.5	55.1	4.8	4.7
		NCS0478	06°28'27''S 76°00'37''W	70514138	8.3	74.3	24	50.6	57.3	5.0	4.9
		NCS0479	06°28'27''S 76°00'37''W	57097304	7.7	73.3	20	48.1	54.4	4.7	4.7
		NCS2554	06°27'41''S 76°20'31''W	51391495	7.5	75.1	18	46.6	52.6	4.6	4.4
		NCS2555	06°27'41''S 76°20'31''W	66232416	7.8	74.7	23	49.7	56.4	4.8	4.7
	emma (rayed)	GS020	06°10'55''S 76°14'50''W	62389463	8.4	75.7	18	44.4	50.3	4.4	4.4
		NCS1671	06°10'55''S 76°14'50''W	67708573	7.5	74.1	24	50.5	56.3	5.0	5.0
		NCS1672	06°10'55''S 76°14'50''W	60859534	7.8	73.9	21	49.0	54.8	4.9	4.8
		NCS1673	06°10'55''S 76°14'50''W	65914675	7.6	74.4	23	50.0	55.8	5.0	5.0
		NCS1674	06°10'55''S 76°14'50''W	60470606	7.9	74.2	22	49.3	55.5	5.0	4.9
		NCS1675	06°10'55''S 76°14'50''W	43527879	8.3	75.5	15	39.1	44.1	4.3	4.2
French Guiana	hydara (postman)	NCS1179	04°42'13''N 52°18'13''W	47188142	8.3	75.2	17	44.9	52.2	4.2	4.0
		NCS1979	04°34'18''N 52°13'24''W	53857631	8.4	75.4	19	48.0	55.7	4.4	4.2
		NCS2080	04°36'28''N 52°16'21''W	52696592	8.3	75.4	20	48.2	56.1	4.4	4.0
		NCS2211	04°32'50''N 52°10'13''W	61935440	8.2	74.8	22	51.2	59.0	4.6	4.2
		NCS2217	04°32'40''N 52°09'09''W	69615030	8.3	75.0	25	52.4	60.1	4.7	4.3
		NCS2581	04°47'48''N 52°19'28''W	87489610	8.0	76.0	29	53.2	61.0	4.8	4.5
		NCS2609	04°47'48''N 52°19'28''W	72210232	8.5	77.2	25	51.4	59.0	4.6	4.3

*SNPs are variation relative to the reference genome

Table S6 (cont.)

hybrid zone	race (phenotype)	sample	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	positions genotyped (%)		SNPs* per genotyped position (%)	
								all reference	*D* interval	all reference	*D* interval
French Guiana (cont.)	erato (rayed)	NCS2005	04°38'19''N 52°18'06''W	64612926	8.6	75.1	21	48.8	54.9	4.6	4.5
		NCS2012	04°38'19''N 52°18'06''W	73271811	8.7	76.4	22	49.4	55.9	4.5	4.4
		NCS2020	04°35'06''N 52°14'44''W	97083432	8.0	75.8	32	53.8	60.2	4.9	4.8
		NCS2023	04°38'19''N 52°18'06''W	76874048	8.3	76.5	23	50.9	57.0	4.6	4.5
		NCS2025	04°35'06''N 52°14'44''W	64208302	8.6	76.0	20	48.1	54.2	4.5	4.4
		NCS2556	04°37'19''N 52°22'34''W	107961988	8.0	72.0	35	55.6	62.2	5.5	5.4
Ecuador	notabilis (postman)	BC0410	01°23'57''S 78°10'52''W	78516169	7.7	75.7	27	54.0	60.8	4.6	4.4
		NOT01	01°23'57''S 78°10'52''W	56434329	8.3	73.9	18	47.8	54.4	4.1	3.8
		NOT02	01°23'57''S 78°10'52''W	58901620	7.8	73.6	18	48.9	55.8	4.3	4.0
		NOT03	01°23'57''S 78°10'52''W	64868484	8.3	73.5	21	50.9	58.1	4.3	4.0
		NOT04	01°23'57''S 78°10'52''W	59804065	8.3	73.7	20	49.9	56.9	4.3	4.1
	lattivita (rayed)	BC0411	01°05'54''S 77°35'02''W	82234945	7.5	76.0	27	51.9	58.4	4.9	4.8
		LAT01	01°05'54''S 77°35'02''W	55007275	7.8	75.4	18	44.3	50.0	4.4	4.4
		LAT02	01°05'54''S 77°35'02''W	70156062	8.1	75.9	22	48.2	54.2	4.6	4.6
		LAT03	01°05'54''S 77°35'02''W	80058495	8.5	76.5	26	50.4	56.7	4.8	4.8
		LAT04	00°42'45''S 77°44'26''W	84018273	8.4	75.5	25	51.3	57.5	4.8	4.7
Panama	petiverana (postman)	ED3	09°07'46''N 79°42'55''W	79848146	8.5	77.3	30	55.6	60.1	3.6	3.9
		ED4	09°07'46''N 79°42'55''W	77997401	8.4	77.2	29	54.5	59.2	3.5	3.7
		ED5	09°07'46''N 79°42'55''W	60922100	8.9	76.8	23	50.8	55.6	3.4	3.6
		ED6	09°07'46''N 79°42'55''W	72039988	8.7	77.0	28	54.1	58.5	3.5	3.7
		STRI0033	09°09'09''N 78°41'23''W	50606981	9.9	67.4	21	52.5	56.8	4.5	3.8
	hydara (postman)	STRI0039	09°09'09''N 78°41'23''W	53723260	8.8	76.5	21	53.3	59.1	3.7	3.8
		STRI0040	09°09'09''N 78°41'23''W	54985081	8.7	76.6	22	54.9	60.3	3.7	3.7
		STRI0042	09°09'09''N 78°41'23''W	55879081	9.1	76.5	21	53.6	59.3	3.8	3.8

Table S7: Samples and sequencing for *H. melpomene* and *H. timareta*

hybrid zone	species (phenotype)	sample ID	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	*B/D* positions genotyped (%)	*B/D* SNPs* per genotyped position (%)
Colombia	*H. melpomene*	HMCS25	4°12'48"N 73°47'70"W	42161297	79.5	93.9	22	74.6	1.8
	melpomene	HMCS27	5°37'01"N 72°18'00"W	66272922	79.4	94.9	34	88.4	1.8
	(postman)	STRI006	5°37'01"N 72°18'00"W	63418043	76.6	93.5	26	81.2	1.7
	H. melpomene	HMCS21	1°48'49"N 75°40'07"W	58085997	75.0	92.1	25	73.7	2.6
	malleti	HMCS22	1°36'35"N 75°40'01"W	52027258	75.1	91.9	23	68.1	2.5
	(rayed)	HMCS24	1°45'02"N 75°37'55"W	44144209	75.6	91.7	19	57.0	2.3
Peru	*H. melpomene*	09-332	see Nadeau et al. 2012					78.6	2.5
	amaryllis	09-333	see Nadeau et al. 2012					79.0	2.5
	(postman)	09-79	see Nadeau et al. 2012					79.6	2.5
		09-75	see Nadeau et al. 2012					75.3	2.4
	H. melpomene	09-246	see Nadeau et al. 2012					68.2	2.6
	aglaope	09-267	see Nadeau et al. 2012					76.8	2.8
	(rayed)	09-268	see Nadeau et al. 2012					76.4	2.8
		09-357	see Nadeau et al. 2012					73.1	2.6
	H. timareta	09-313	6°27'11"S 76°17'19"W	59607967					
	(aligned full reference)				67.4	92.0	22	84.7	2.2
	(aligned partial reference)				4.9	71.7	28	76.2	2.3

*SNPs are variation relative to the reference genome

S4. Population Genetic Analyses between Divergent Races

S4.1 Population Genetic Methods

We used a number of population genetic analyses to identify putative functional regions. We examined signatures of selection, including increased genomic divergence between divergent color pattern races, and genotype by phenotype association to highlight regions showing patterns consistent with strong selection acting on functional variation.

Signatures of selection

We examined genomic divergence between pairs of *H. erato* color pattern races at each of four hybrid zone independently and across all three postman/rayed hybrid zones combined. To analyze each hybrid zone independently, we calculated sliding window population differentiation using a method that uses diploid data and models populations as random effects, to account for both statistical and genetic sampling processes ($\hat{\theta}$, Weir 1996). The model makes no simplifying assumptions regarding sample sizes or number of populations (Weir and Cockerham 1984). Calculations were done using a custom Perl script that implemented the Bio::PopGen::PopStats module from BioPerl (www.bioperl.org). To examine genomic divergence between the red phenotypes across the three postman/rayed *H. erato* hybrid zones combined, while accounting for the geographic structure of the populations, we estimated differentiation in a three-level hierarchy method ($\hat{\theta}_S$, Weir 1996). For level one, the populations, we examined the three hybrid zones that showed variation in the red phenotype—Peru, French Guiana, and Ecuador. For level two, the subpopulations, we examined the two color pattern races at each hybrid zone—the postman and the rayed. For level three, the individuals, we examined five to eight individuals per subpopulation. We calculated the sliding window subpopulation differentiation ($\hat{\theta}_S$) using a custom BioPerl module. For all comparisons, we calculated divergence at a position only if at least 75% of the individuals were genotyped for each phenotype. We evaluated 15 kb sliding windows at 5 kb steps across the genomic intervals and required a window to have divergence calculated for at least 20% of the positions in the window. We calculated a baseline level of divergence for each comparison as the level of divergence observed across intervals unlinked to color pattern (*H. erato*—three unlinked BACs, *H. melpomene*—38 unlinked scaffolds).

We calculated sliding window values for the proportion of segregating sites and heterozygosity to look for signatures of selection in *H. erato*. A segregating site was defined as having more than one allele in the population. The proportion of segregating sites was the total number of segregating sites per window divided by the total number of sites examined in the window. We obtained the proportion of heterozygotes by summing the number of heterozygote individuals at each position in the window and dividing that by the sum of the number of individuals genotyped at that position. We calculated baseline values from the three contigs unlinked to color pattern. We calculated estimates of these parameters for a genomic position only if at least 75% of individuals were genotyped and then examined 15 kb windows with a 5 kb step size. We required a window to have parameters estimated for at least 20% of the positions in the window.

Genotype by phenotype analyses

We estimated genotype by phenotype association at each *H. erato* hybrid zone independently, comparing the two color pattern phenotypes that occur in each hybrid zone. We also examined association with red phenotype across all four *H. erato* hybrid zones combined, by assigning all individuals to one of the two major red phenotypes—the postman or the rayed. We estimated association at each biallelic SNP using a two tailed Fisher's exact test, based on allele counts. Positions were excluded if less than 75% of individuals were genotyped for each phenotype.

Population genetic analyses in *H. melpomene*

We also assessed divergence and association in the *H. melpomene* hybrid zones in Peru and Colombia, which both consists of the two major red phenotypes—the postman and the rayed. We calculated sliding window subpopulation differentiation and genotype by phenotype association as described above. Additionally, we compared the positions of fixed SNPs between *H. erato* and *H. melpomene* to determine if any shared fixed SNPs existed. For each fixed SNP in *H. erato*, we attempted to identify an orthologous SNP in *H. melpomene* by manually inspecting the mVista LAGAN alignment between the reference sequences for the two species. If we were able to identify an orthologous SNP, we then compared the genotype calls for *H. erato* and *H. melpomene* individuals to determine if the SNP was associated with phenotype in both species.

S4.2 Population Genetic Results

See main text for population genetic results.

S5. Linkage Disequilibrium (LD) and Haplotype Structure

S5.1 LD and Haplotype Methods

We explored linkage disequilibrium (LD) and haplotype structure across the *D* interval, and regions unlinked to color pattern, in the Peruvian hybrid zone. We focused on a single hybrid zone because we wanted to remove the influence of geography and we chose Peru because it had the largest sample size. The data included all biallelic SNPs with at least 75% of individuals genotyped. We calculated correlations (r^2) between all pairwise SNPs using PLINK (Purcell et al. 2007), which for unphased data is based on genotype allele counts. To understand how LD decays with the distance between SNPs, we averaged the correlations for all pairwise SNPs from 100 bp bins of distance.

We estimated haplotypes from the Peruvian hybrid zone across a 100 kb window of the *D* interval (500-600 kb) containing the 65 kb peak of divergence and flanking regions using fastPHASE v1.2 (Scheet and Stephens 2006). We filtered bialleleic SNPs across this 100 kb region to remove sites that had genotypes from less than 75% of the individuals of each race, resulting in 3227 SNPs. Haplotypes were clustered during phase estimation into two clusters (K=2) and the proportion of rayed and postman individuals assigned to each cluster at each SNP was determined. We used Haploscope (San Lucas et al. 2012) to visualize regions where the two races had fixed haplotype block differences and where individuals from both races shared the same haplotypes. Using the haplotype estimations from fastPHASE, for each SNP Haploscope

visualizes the portion of individuals from a race (light vs. dark) assigned to each cluster (red vs. grey) across the 100 kb region (San Lucas et al. 2012).

S5.2 LD and Haplotype Results

See main text for LD and haplotype results.

S6. Phylogenetic Analyses of Evolutionary History

S6.1 Phylogenetic Methods

We constructed phylogenetic trees across sliding windows in the *D* interval, sampling 15 kb of sequence every 5 kb. For each window, we tested the log likelihood of the data with two alternative trees: the geographic tree assumes samples cluster by geographic hybrid zone and the color based tree groups races with a similar color pattern (rayed or postman) in a monophyletic clade (Figure 5). In each case, races are assumed to be monophyletic so that hypotheses of racial structure are equivalent. Neither geographic regions nor similarly colored races were resolved relative to one another, to avoid the influence of other topological hypotheses on the results. Likelihood values were calculated for each interval and tree topology using scripts in PAUP* 4b10 (Swofford 2002), using a GTR + G model inferred for the interval as a whole using Modeltest v3.7 (Posada and Crandall 1998). In addition to calculating likelihoods, we constructed neighbor-joining trees across these sliding windows in PAUP* to infer where in the interval lineages were monophyletic by color phenotype.

To summarize variation in phylogenetic topology across the interval we constrained division of the interval into the five most distinct topologies using the MDL method (Ané 2011). Default likelihood penalties for this method support a different topology for every block of 500 consecutive SNPs assessed. To divide the region more broadly, we raised the likelihood score penalty until five clusters of SNP blocks were reached. Tree topologies for each of these five regions of the interval were constructed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) run on CIPRES Science Gateway (Miller et al. 2010). Analyses involved 3 runs for 3 million generations each, sampling every 500 generations and removing 33% burn-in and runs that did not converge (as assessed in MrBayes and Tracer v1.5 (Rambaut and Drummond 2007)). Models were assigned using MrModeltest v2.3 (Nylander 2004) and included the GTR model for the 2nd and 3rd regions of the interval and GTR+G for the remaining regions.

In addition to these phylogenies, to infer a "best" tree of color pattern history, a phylogenetic tree was constructed for the 515-580 kb region across the peak of population differentiation. This tree was constructed under the same parameters in MrBayes (model = GTR) using only SNPs with low missing data, including at least 75% coverage of individuals within each red phenotype (1419 SNPs). A general phylogeny was also constructed across three genomic regions unlinked to color pattern using variable sites with at least 75% coverage of individuals under the same Bayesian methods (model = GTR + G, 3534 SNPs). For these "best" color-linked and color-unlinked datasets, we also reconstructed unrooted neighbor-net splits tree networks using SplitsTree v4.8 (Huson and Bryant 2006) and pairwise distances. Unlike most other phylogenetic analyses, which treat polymorphisms as ambiguities ("W" as "A or T"), in this analysis we were able to treat characters additively as "averages" ("W" as "A and T").

When treated as averages, sites where all individuals are heterozygotes are informative, thus more sites were retained as variable for this analysis (3440). Treating sites as averages should more accurately reflect the history of these characters which are by nature additive: two heterozygotes are more similar to each other than they are to homozygotes of either allele and in the additive model, heterozygotes are treated with 50% similarity to homozygotes, rather than as identical. These phylogenetic networks have the additional advantage of graphically representing areas and degrees of character conflict in phylogenetic construction, brought on through hybridization and recombination, ancestral sorting, or homoplasy.

To test whether shared color patterns between the mimics could result from a common origin, we also performed phylogenetic analyses combining *H. erato* and *H. melpomene* sequences along this interval. We focused on regions of high conservation between *H. erato* and *H. melpomene* (>80% conservation in a 500 bp window) from our mVista alignment. Across the 450 to 750 kb interval, we found 71 highly conserved regions that were relatively evenly distributed across the region and ranged in size from 430 to 3857 bp. For each conserved region, we used ClustalW2 (Larkin et al. 2007) to align sequences from all 45 *H. erato* individuals (Table S6) and 14 *H. melpomene* individuals (Table S7). We constructed neighbor-joining trees from pairwise distances of taxa in each these fragments and examined the resulting trees for species monophyly.

To infer a "best" *D* locus tree of *H. melpomene* and *H. erato* combined, we inferred the history in the peak of association from 515-580 kb in *H. erato* after further filtering SNPs from the regions of high conservation. This included first manually editing the alignments by removing regions of highly ambiguous alignment and correcting obvious misalignments. We then removed invariant sites and sites with more than 25% missing data. The resulting 1134 SNPs were concatenated and used for a Bayesian analysis using all the same parameters as listed above, including a GTR model inferred independently for this dataset in MrModeltest. We characterized SNPs across the interval by their patterns of fixation with respect to species and phenotype.

S6.2 Phylogenetic Results

To infer the optimal history of color pattern diversification in *H. erato* we constructed a Bayesian tree and a network-based tree of the 65 kb region that showed the strongest divergence and color pattern association. These trees support a single origin of the rayed color pattern, clustering rayed phenotypes separate from non-rayed phenotypes (Figure S6A). Trees based on SNPs from color-pattern unlinked regions clustered largely by hybrid zone (Figure S6B). Branch lengths across topologies show a signature of reduced gene flow, whereby color pattern alleles have reduced gene flow among races and less variation among individuals relative to markers unlinked to color pattern loci.

Comparing the history of this region between the two co-mimics, *H. erato* and *H. melpomene*, is difficult, as aligning non-coding regions is problematic; however, we were able to align regions of conservation between the two species. Of the 71 aligned fragments within the 300 KB window including the association peaks, 69 resulted in complete monophyly of *H. melpomene* with respect to *H. erato*. The remaining two trees had a few individuals admixing between the two in a manner unrelated to phenotype. Examination of the sequence files for

these fragments revealed problems with the automated alignment due to extensive missing data for these taxa.

We focused further analyses on the 65 KB region of highest association. After manual alignment and removal of sites with greater than 25% missing data in this narrowed region, 1164 SNPs were retained. Among these SNPs, 360 were fixed by species, 591 varied only in *H. erato*, 140 varied only in *H. melpomene*, and 73 shared allelic variation between the two species. Of the SNPs with alleles shared between the two species, none of the alleles that were fixed by phenotype within *H. melpomene* (n=18) or *H. erato* (n=1) had a signal that was associated with color phenotype in the opposite species. A phylogenetic analysis of the 1164 SNPs resolves *H. erato* and *H. melpomene* as separate lineages with high support, while resolving races by phenotype within each species (Figure S5). Results from these data thus support an independent origin of red patterns within each species.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Andrews S. 2011. FastQC v0.8.0. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biology and Evolution* **3**: 246–258.

Beltrán M, Jiggins CD, Brower AV, Bermingham E, and Mallet J. 2007. Do pollen feeding, pupal-mating and larval gregariousness have a single origin in *Heliconius* butterflies? inferences from multilocus DNA sequence data. *Biological Journal of the Linnean Society* **92**: 221–239.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**: 365–370.

Broad Institute. 2009. Picard. http://picard.sourceforge.net.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, Green ED, Sidow A, and Batzoglou S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**: 721–731.

Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–930.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**: 677–681.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, and Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676.

Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, Papa R, Ferguson L, Joron M, ffrench Constant RH, et al. 2010. Genomic hotspots for adaptation: The population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genetics* **6**: e1000796. doi:10.1371/journal.pgen.1000796.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**: 491–498.

Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, Wu Y, Wang J, Mita K, Xiang Z, et al. 2010. SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Research* **38**: D453–D456.

Flanagan NS, Tobler A, Davison A, Pybus OG, Kapan DD, Planas S, Linares M, Heckel D, and McMillan WO. 2004. Historical demography of Müllerian mimicry in the neotropical *Heliconius* butterflies. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 9704–9709.

Gordon A. 2010. FASTX Toolkit v0.0.13.1. http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–98.

Holt C and Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491. doi:10.1186/1471–2105–12–491.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al. 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* **40**: D306–D312.

Huson DH and Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**: 254–267.

Illumina. 2008. *Preparing Samples for Sequencing of mRNA*. Illumina.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59. doi:10.1186/1471–2105–5–59.

Langmead B, Trapnell C, Pop M, and Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25. doi:10.1186/gb–2009–10–3–r25.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Li H and Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**: 1754–60.

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

Mallet J. 1986. Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity* **56**: 191–202.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

McQuilton P, St Pierre SE, Thurmond J, and the FlyBase Consortium. 2012. FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Research* **40**: D706–D714.

Miller JR, Koren S, and Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315–327.

Munoz-Torres MC, Reese JT, and Sundaram JP. 2011. *Bee gene model annotation using Apollo*. Elsik Computational Genomics Laboratory.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**: 343–353.

Nylander JAA. 2004. Mrmodeltest v2. Program distributed by the author.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, and Sonnhammer ELL. 2010. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* **38**: D196–D203.

Papanicolaou A, Stierli R, ffrench Constant R, and Heckel D. 2009. Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics* **10**: 447. doi:10.1186/1471–2105–10–447.

Pohl N, Sison-Mangus M, Yee E, Liswi S, and Briscoe A. 2009. Impact of duplicate gene copies on phylogenetic analysis and divergence time estimates in butterflies. *BMC Evolutionary Biology* **9**: 99. doi:10.1186/1471–2148–9–99.

Pool JE, Hellmann I, Jensen JD, and Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Research* **20**: 291–300.

Posada D and Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Pruitt KD, Tatusova T, and Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**: D61–D65.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559 – 575.

Rambaut A and Drummond AJ. 2007. Tracer v1.4. http://beast.bio.ed.ac.uk/Tracer.

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, Jiggins CD, Chamberlain NL, Kronforst MR, Chen R, et al. 2011. optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**: 1137–1141.

Ronquist F and Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

San Lucas FA, Rosenberg NA, and P S. 2012. Haploscope: a tool for the graphical display of haplotype structure in populations. *Genetic Epidemiology* **36**: 17–21.

Scheet P and Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* **78**: 629 – 644.

Schug J and Overton GC. 1997. TESS: Transcription element search software on the WWW. In *Technical Report CBIL-TR-1997-1001-v0.0*. Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania.

Slater G and Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31. doi:10.1186/1471–2105–6–31.

Smit AFA, Hubley R, and Green P. 2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, and Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**: W435–W439.

Suzek BE, Huang H, McGarvey P, Mazumder R, and Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282–1288.

Swofford DL. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, 4th edition. Sinauer Associates, Sunderland, Massachusetts.

Trapnell C, Pachter L, and Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511–515.

Vergara IA and Chen N. 2009. Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Current Protocols in Bioinformatics* pp. 1–18.

Weir BS. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, Massachusetts.

Weir BS and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.