Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*

Nicola J. Nadeau,1,2 Mayté Ruiz,3 Patricio Salazar,1,4 Brian Counterman,5 Jose Alejandro Medina,6 Humberto Ortiz-Zuazaga,6,7 Anna Morrison,1 W. Owen McMillan,8 Chris D. Jiggins,1,8 and Riccardo Papa3

1Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom; 2Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, United Kingdom; 3Department of Biology and Center for Applied Tropical Ecology and Conservation, University of Puerto Rico, Rio Piedras, San Juan, Puerto Rico 00921; 4Centro de Investigación en Biodiversidad y Cambio Climático (BioCamb), Universidad Tecnológica Indoamérica, Quito, Ecuador; 5Department of Biology, Mississippi State University, Mississippi 39762, USA; 6High Performance Computing Facility, University of Puerto Rico, San Juan, Puerto Rico, 00921; 7Department of Computer Science, University of Puerto Rico, Rio Piedras, San Juan, Puerto Rico 00921; 8Smithsonian Tropical Research Institute, Apartado 0843-03092, Balboa, Ancón, Panama

Hybrid zones can be valuable tools for studying evolution and identifying genomic regions responsible for adaptive divergence and underlying phenotypic variation. Hybrid zones between subspecies of *Heliconius* butterflies can be very narrow and are maintained by strong selection acting on color pattern. The comimetic species, *H. erato* and *H. melpomene*, have parallel hybrid zones in which both species undergo a change from one color pattern form to another. We use restriction-associated DNA sequencing to obtain several thousand genome-wide sequence markers and use these to analyze patterns of population divergence across two pairs of parallel hybrid zones in Peru and Ecuador. We compare two approaches for analysis of this type of data—alignment to a reference genome and de novo assembly—and find that alignment gives the best results for species both closely (*H. melpomene*) and distantly (*H. erato*, 15% divergent) related to the reference sequence. Our results confirm that the color pattern controlling loci account for the majority of divergent regions across the genome, but we also detect other divergent regions apparently unlinked to color pattern differences. We also use association mapping to identify previously unmapped color pattern loci, in particular the *Ro* locus. Finally, we identify a new cryptic population of *H. timareta* in Ecuador, which occurs at relatively low altitude and is mimetic with *H. melpomene malletii*.

[Supplemental material is available for this article.]

Natural hybrid zones occur where divergent forms meet, mate, and hybridize. Narrow hybrid zones can be maintained by strong selection that prevents mixing or favors particular forms in particular areas (Barton and Hewitt 1985). Studies of hybrid zones have provided many insights into the origins of diversity and the process of speciation (Mallet et al. 1990; Harrison 1993; Kawakami and Butlin 2001). High-throughput sequencing technologies now provide the opportunity for hybrid zones to fully meet their potential as windows into the evolutionary process by allowing us to move beyond studies of neutral variation at a handful of loci and identify the genetic loci under selection (Rieseberg and Buerkle 2002; Gompert et al. 2012; Crawford and Nielsen 2013).

Butterflies of the Neotropical genus *Heliconius* are extremely diverse in their wing color patterns and combine within species diversity with convergence among species in wing phenotypes. Their bright wing patterns are used as aposematic warnings to predators and are under positive frequency-dependent selection favoring common color patterns that predators learn to avoid. This strong selection also maintains narrow hybrid zones between subspecies with different patterns (Benson 1972; Mallet and Barton 1989a; Kapan 2001; Langham 2004). In addition, frequency-dependent selection leads to Müllerian mimicry between many distinct species (Müller 1879). For instance, *H. erato* and *H. melpomene* are two distantly related species that diverged ~15–20 million years ago, but have converged on common color patterns across most of the Neotropics. Divergent races of both species meet in parallel hybrid zones (Fig. 1). Evidence suggests that convergent color patterns in these two species have evolved independently (Hines et al. 2011; Supple et al. 2013). It has also been suggested that *H. erato* is more ancient and *H. melpomene* diversified more recently to mimic the *H. erato* forms (Brower 1996; Flanagan et al. 2004; Quek et al. 2010). Nevertheless, it appears that the same handful of genetic loci are responsible for producing most of the color pattern variation in both species (Joron et al. 2006; Baxter et al. 2008; Reed et al. 2011; Martin et al. 2012). This pattern of parallel adaptive radiation makes *Heliconius* an excellent system in which to address the predictability of the evolutionary process and the extent to which particular genes are re-used when evolving the same phenotypes (Papa et al. 2008a; Nadeau and Jiggins 2010).

In this study, we use high-resolution genome scans to investigate patterns of divergence across two pairs of parallel hybrid zones in Peru and Ecuador. These occur between subspecies with...
different wing color patterns in both *H. erato* and *H. melpomene* (Fig. 1). In both regions, the clines in color pattern alleles between species are highly coincident (Mallet et al. 1990; Salazar 2012). The two hybrid zones in Peru have been the focus of several previous studies, whereas those in Ecuador have been less well studied. In Peru, strong natural selection has been shown to maintain color pattern differences (Mallet and Barton 1989a) and loci controlling color patterns show enhanced divergence (Baxter et al. 2010; Counterman et al. 2010; Nadeau et al. 2013; Martin et al. 2013; Supple et al. 2013). However, we still lack a complete picture of how many loci are divergent between subspecies and the extent to which the genomic architecture of divergence is the same between mimetic species.

Extensive genetic mapping using experimental crosses between different color pattern forms has identified the chromosomal regions responsible for color pattern variation (Sheppard et al. 1985; Joron et al. 2006; Baxter et al. 2008; Papa et al. 2013). Three major clusters of loci control most of the color pattern variation observed in both species. The tightly linked *B* and *D* loci on chromosome 18 in *H. melpomene* control the red forewing band, and the red/orange hindwing rays and proximal "dennis" patches on both wings, respectively. These loci are homologous to the *D* locus in *H. erato* (Baxter et al. 2008) and appear to be *cis* regulatory elements of the *optix* gene (Reed et al. 2011; Supple et al. 2013). The *Ac* and *Sd* loci, in *H. melpomene* and *H. erato*, respectively, control the shape of the forewing band via regulation of the *WntA* gene on chromosome 10 (Martin et al. 2012). The presence of most yellow and white elements on the wing is largely controlled by three tightly linked loci, *Yb*, *Sb*, and *N*, on chromosome 15 in *H. melpomene* (Ferguson et al. 2010), which are homologous to the *Cr* locus in *H. erato* (Supple et al. 2013). Quantitative trait locus (QTL) mapping has identified other loci of minor effect, including at least seven additional QTLs.
in *H. erato* (Papa et al. 2013), and QTL in *H. melpomene* on chromosomes 2, 7, and 13 that affect forewing band shape (Baxter et al. 2008). In some cases, mapping studies have been followed up by population genetic studies of the mapped intervals across natural hybrid zones, where many generations of backcrossing have led to narrow regions of association, permitting fine scale mapping (Baxter et al. 2010; Counterman et al. 2010; Nadeau et al. 2012; Supple et al. 2013). High-throughput sequencing technologies now provide the feasibility to generate a high density of genomic markers to identify the narrow QTL present in these hybrid zones without the need to perform controlled laboratory crosses (Crawford and Nielsen 2013). Here we test this approach, using a system in which some of the loci responsible for phenotypic differences are known.

The Peru and Ecuador hybrid zones occur across altitudinal gradients (Fig. 2A). Therefore, it is possible that traits other than color pattern may also be differentiated by altitudinal selection; for example, traits related to temperature or changes in larval host plants. Such selection on additional regions of the genome could help to stabilize the geographic location of the hybrid zone (Barton and Hewitt 1985; Mallet and Barton 1989b; Mallet 2010; Bierne et al. 2011). Therefore another important question that we will address is whether there are divergent regions of the genome that are not controlling color pattern. These might be candidates for loci controlling other aspects of ecological adaptation.

In this study we use restriction-associated DNA (RAD) sequencing (Baird et al. 2008) to determine, for the first time:

1. If association mapping in these hybrid zones can identify known and novel loci underlying phenotypic variation;
2. How much of the genome is differentiated and under divergent selection between subspecies;
3. How much of this differentiation is due to loci controlling color pattern variation;
4. If the same regions are divergent between co-mimetic species.

Although previous studies have touched on questions 2 and 3 (Kronforst 2013; Martin et al. 2013), here we focus on divergence at the subspecies level where hybridization is frequent, rather than between occasionally hybridizing species. Compared to the study by Martin et al. (2013), we explored additional hybrid zones (Ecuador and species *H. erato*) using larger sample sizes, which allowed more robust tests to identify genomic regions under divergent selection. We also investigate the advantages and limitations of alignment and assembly methods when only a single reference genome is available. We compare two widely used approaches: de novo assembly of just the restriction-associated reads, using the program Stacks (Catchen et al. 2011), versus alignment of paired-end reads to the reference *H. melpomene* genome.

**Results**

**Summary of the data and comparison of alignment and assembly techniques**

We sequenced a total of 129 individuals of *H. erato* and *H. melpomene* from the four hybrid zones in Peru and Ecuador, including a small number of additional individuals from across the range of *H. erato*. Using restriction-associated DNA sequencing (RAD-seq), we obtained a total of 1496M 150 bp paired-end reads from the hybrid zone individuals, and an additional 115M 100 bp paired-end reads from the other *H. erato* populations and outgroups. In our analyses, we also include data from additional *H. melpomene* populations and outgroups (*H. cydno, H. timareta,* and *H. hecale*) already published in a previous study (Nadeau et al. 2013).

Our reference genome for *H. melpomene* is highly divergent from *H. erato*. Nonetheless, for both species, alignment of reads to the *H. melpomene* reference sequence yielded more usable data when compared to de novo assembly carried out independently within each species. De novo assembly produced more bases in assembled contigs (Table 1), but only ~2% of contigs assembled in the *H. erato* populations were present in more than 10 individuals, with the figure being ~7% in *H. melpomene*. By comparison, when the same data (plus the paired-end reads) were aligned to our reference sequence, ~38% of aligned bases were found in more than 10 individuals in *H. erato* and >50% in *H. melpomene*. We hypothesized that high levels of within-population variation led to homologous reads being separated into distinct contigs in the de novo assembly. We could confirm that this was the case for one region of the *H. erato* genome for which a high quality reference sequence is available (Supple et al. 2013). Across 960 kb at the *D* color pattern locus, we observed that RAD-seq reads that were highly divergent between subspecies could be aligned to homologous positions in the reference but were assembled into separate contigs in the de novo assembly. Overall, we also found a higher frequency of single nucleotide polymorphisms (SNPs) in the reference alignments than the de novo assemblies (Table 1). These SNPs were defined as sites that were polymorphic within the sampled populations and so are not inflated by fixed differences from the reference genome.

As expected, given that *H. erato* is ~15% divergent from *H. melpomene* in the aligned data, fewer *H. erato* reads aligned to the *H. melpomene* genome as compared to those from *H. melpomene*, leading to fewer confidently called bases. Nevertheless, the use of the reference *H. melpomene* genome for aligning the *H. erato* reads resulted in more bases being called across multiple individuals and around 10× more SNPs identified when compared to the de novo assembly approach. In addition, the gaps between aligned RAD loci were similar across both species (Table 1), indicating that the reduced number of bases is not due to fewer RAD loci aligning but to fewer confidently called bases at each RAD locus. The power to detect loci under selection or responsible for phenotypic variation should therefore be similar in both species or slightly reduced in *H. erato* due to its larger genome (Tobler et al. 2004). Nevertheless, much of the additional genomic sequence in *H. erato* is likely to be repetitive DNA (Papa et al. 2008b), which would be difficult to align and score variants in, even if a complete reference was available. In summary, it seems that the aligned data should give the most power to detect divergent regions and phenotypic associations for both species. However, we performed outlier and association analyses using the output of both approaches for comparison. It is also possible that the de novo assembly might detect divergent regions important in adaptation that could not be aligned to the *H. melpomene* reference.

**Phylogenetics and population structure**

Using the reference aligned sequence data, we constructed maximum likelihood phylogenies for the *H. melpomene* and *H. erato* clades, including individuals from additional populations and outgroup taxa (Fig. 1). This revealed remarkably similar patterns of divergence between co-occurring, co-mimetic subspecies in both groups. Population divergence in *H. erato* is thought to be deeper than that in *H. melpomene* (Flanagan et al. 2004, but see Cuthill and Charleston 2012), but this was not evident in our tree as branch
Figure 2. Population structure at each of the hybrid zones using the reference aligned data. (A) Sampling locations with altitude in meters, sample size in parentheses, and pie charts of the proportion of individuals of each type sampled from each site. Colors are the same as in Figure 1, except black indicates *H. timoreta* in Ecuador. (B) Structure analysis with $k=2$ (*H. timoreta* individuals excluded). Each individual is shown as a horizontal bar with the allelic contribution from population 1 (gray) and population 2 (black). (C) Principal components analysis. (D) Distribution of $F_{ST}$ values from BayeScan.
### Table 1. Summary statistics from alignment and assembly approaches

<table>
<thead>
<tr>
<th>Strain</th>
<th>n</th>
<th>Millions of reads (mean ±SD)</th>
<th>Bases covered (×10⁶)</th>
<th>Bases covered in ≥10 inds (×10⁶)</th>
<th>SNPs used in outlier analysis (×10³)</th>
<th>Mean $F_{ST}$</th>
<th>Outliers</th>
<th>Significant phenotypic associations</th>
<th>Mean gap between RAD loci (kb)</th>
<th>Maximum gap between RAD loci (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>De novo assembly with Stacks: single end reads</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H. erato</td>
<td>Peru</td>
<td>27</td>
<td>8.0 ± 2.2</td>
<td>166</td>
<td>2.8</td>
<td>37</td>
<td>0.0280</td>
<td>22</td>
<td>10</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Ecuador</td>
<td>30</td>
<td>9.2 ± 2.5</td>
<td>149</td>
<td>3.3</td>
<td>31</td>
<td>0.0568</td>
<td>0</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>H. melpomene</td>
<td>Peru</td>
<td>30</td>
<td>7.0 ± 2.4</td>
<td>61</td>
<td>4.3</td>
<td>57</td>
<td>0.0145</td>
<td>23</td>
<td>8</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Ecuador</td>
<td>22</td>
<td>7.8 ± 1.4</td>
<td>45</td>
<td>3.5</td>
<td>43</td>
<td>0.0310</td>
<td>5</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Aligned to H. melpomene reference: paired end with Stampy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H. erato</td>
<td>Peru</td>
<td>27</td>
<td>11.3 ± 3.2</td>
<td>11</td>
<td>4.2</td>
<td>373</td>
<td>0.0142</td>
<td>19</td>
<td>28</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>Ecuador</td>
<td>30</td>
<td>11.9 ± 3.2</td>
<td>13</td>
<td>5.1</td>
<td>337</td>
<td>0.0316</td>
<td>56</td>
<td>15</td>
<td>9.1</td>
</tr>
<tr>
<td>H. melpomene</td>
<td>Peru</td>
<td>30</td>
<td>10.9 ± 3.8</td>
<td>28</td>
<td>14.4</td>
<td>860</td>
<td>0.0112</td>
<td>235</td>
<td>91</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td>Ecuador</td>
<td>22</td>
<td>10.7 ± 1.9</td>
<td>23</td>
<td>15.6</td>
<td>788</td>
<td>0.0299</td>
<td>179</td>
<td>14</td>
<td>9.5</td>
</tr>
</tbody>
</table>
lengths were very similar between the two species. This may be due to the lower quality of alignments for H. erato, with the H. erato tree based on about a third as many sites as that for H. melpomene. These sites in H. erato are likely to be more conserved, resulting in some compression of the tree topology.

The most striking finding from the phylogenetic reconstruction was that eight of the presumed H. melpomene individuals from Ecuador were strongly supported as clustering within the H. timareta clade (Fig. 1). All these individuals had a H. melpomene malletii-like phenotype with the exception of one individual which had been characterized as a possible hybrid due to a large and rounded yellow forewing band, but was otherwise H. m. malletti-like. This finding was surprising because although populations of H. timareta mimetic with H. m. malletti have previously been described in Colombia (Giraldo et al. 2008) and Northern Peru (Lamas 1997), they are all found in highland areas above ~1000 m. Similar populations are not known from lowland sites anywhere in the range. To compare our individuals to these and other populations, we also directly sequenced part of the mitochondrial COI gene that overlaps with the region sequenced in previous studies (Giraldo et al. 2008; Metz Jr. et al. 2013). Our phylogeny based on these sequences also robustly supported these eight individuals as being H. timareta and placed them closer to the highland H. timareta timareta in Ecuador than to H. timareta florencia in Colombia that they resemble phenotypically (Supplemental Fig. 1).

The newly identified H. timareta subspecies was also clearly evident in a principal components analysis (PCA) of the combined H. melpomene, H. timareta, and H. cydno data. The first principal component separated the Peruvian H. melpomene from H. timareta and H. cydno (which were very similar on this axis) (Supplemental Fig. 2). The grouping of the Ecuadorian samples was consistent with the phylogeny, with the same eight individuals clustering with H. timareta. No individuals were intermediate between H. melpomene and H. timareta, indicating that the level of genetic isolation between the two species is similar to elsewhere in their range. This was also confirmed by a Structure analysis of the Ecuador “H. melpomene” population, where a model with two populations had the best fit to the data (posterior probability = 1). Under this model, which allowed for admixture between populations, the H. timareta individuals all had 100% of their allelic contribution from population 1, whereas for H. melpomene the maximum contribution of population 1 to any individual’s genotype was 1.8% (Supplemental Table 1). In summary, we can conclude that these are distinct species with little gene flow between them.

We conducted further analyses of the genetic structure of each of the hybrid zone populations using the reference aligned data, excluding the H. timareta individuals. Overall, these results suggest only very low genetic differentiation between any of the parapatric subspecies. Structure analyses of each population generally showed very little structure and strongest support for only a single population cluster being present. The only exception was the Peruvian H. melpomene, where two population clusters gave the highest posterior probability (P = 1). However, these clusters did not correspond to the two subspecies. The genetic diversity was partitioned such that most individuals were admixed with about a quarter of their allelic variation from population 2, except for two “hybrid” individuals that had pure population 2 genotypes and two other individuals (one “hybrid” and one aglaope) that had almost pure population 1 genotypes (Fig. 2B). PCA revealed very similar patterns, with small groups of hybrid phenotype individuals giving the clearest clusters, which in most cases were also identified by Structure (with K = 2) (Fig. 2C). Three of the populations did reveal some separation of the subspecies at one of the first two principal components, but with a gradual change from one genomic “type” to another. The H. melpomene subspecies in Ecuador were separated by PC1, which explained 10% of the variation in this population. The two H. erato populations both showed some separation by subspecies at PC2, which explained 5.7% and 6.7% of the variation in Peru and Ecuador, respectively. We found very similar results with PCA on the de novo assembled data (Supplemental Fig. 3), suggesting that the underlying genetic signal in both data sets is very similar. The lack of strong differentiation between subspecies was also supported by the FsST distributions (calculated by BayeScan), which gave very low FsST values between subspecies at >99% of the genome, with only a small percentage of SNPs showing high levels of differentiation (Fig. 2D; Supplemental Fig. 3).

**Association mapping of loci responsible for phenotypic variation**

We performed association mapping to identify genetic regions responsible for the phenotypic variation that segregates across each of the hybrid zones. In general, the expected associations were found at the three major loci known to control color pattern variation on chromosomes 10, 15, and 18 (Figs. 3A,D, 4A,D; Table 2; Supplemental Table 2). The majority of SNPs showing significant phenotypic associations fell within or tightly linked to these loci in all populations except in Peruvian H. erato, where only 26% were tightly linked to the known loci (Supplemental Figs. 4–8; Supplemental Table 2).

Independent analyses were performed on both the reference alignments and de novo assemblies of the data. In all populations, more associated SNPs were identified in the alignments than in the de novo assemblies (Fig. 5; Table 1). We used BLASTN (Altschul et al. 1990) to place de novo contigs containing associated SNPs onto the H. melpomene genome, and most could be confidently assigned to a unique locus. There was almost no overlap in the particular SNPs detected in the assembled and aligned data sets (Fig. 5), although in many cases the SNP’s detected were in similar regions (Figs. 3, 4). There was evidence for a higher false positive rate in the de novo data, because the majority of the SNPs that were uniquely significantly associated in these data were present in the aligned data but did not reach significance. This, rather than detection of novel regions, appears to be the main cause of the higher proportion of associations found scattered across the genome and away from known color pattern loci in the de novo data.

**Red color pattern elements and the B and D loci**

Our mapping of red color pattern variation was generally consistent with previous studies (Baxter et al. 2008; Counterman et al. 2010; The Heliconius Genome Consortium 2012) and in almost all populations was mapped to the expected region of chromosome 18 (Table 2; Supplemental Table 2). The only exception was H. melpomene in Ecuador, where SNPs in this region did not reach significance (Fig. 4A). This is likely to be due to the reduced sample size of this population (22) after removing the H. timareta individuals. Color patterns were scored both as independent elements and also using known patterns of segregation to score the predicted genotype at the B/D locus. The red forewing band and red hindwing rays are both dominant traits but are controlled by a single locus (or very tightly linked loci) and inherited in repulsion, meaning that individuals with both traits can be inferred to be heterozygotes (Sheppard et al. 1985; Baxter et al. 2008). This

---

**Population genomics of parallel hybrid zones**

---
genotypic scoring generally gave stronger associations (Figs. 3, 4), although both methods gave some significant associations for at least one of the traits. In all populations, the strongest associations in this region were >60 kb downstream from the optix gene that controls red color pattern (Reed et al. 2011), overlapping the region identified in previous analyses as likely containing the functional regulatory variation (Table 2; Supplemental Fig. 4; Nadeau et al. 2012; Supple et al. 2013).

In several populations we found additional associations with B/D phenotypes on linked chromosome 18 scaffolds (Supplemental Table 2). The furthest from the B/D locus was HE671488 in Peruvian H. melpomene, which is ~2 Mb away. This scaffold was also associated with differences in altitude in this population, which were stronger than the associations with color (Fig. 3A; Table 3; Supplemental Fig. 4). This could suggest that this B/D linked region is responsible for ecological adaptation, although

Figure 3. Association mapping (A,D) and outlier analysis (B,E) for H. melpomene (A–C) and H. erato (D–F) in Peru. Each phenotype used for the association mapping is shown in a different color as illustrated in C and F. For clarity, only the top 20 associated SNPs are shown for each phenotype. The outlier analysis results show $F_{ST}$ values for all SNPs, with significant outliers shown in red. Results from the de novo assembled data are shown as crosses (and in orange for the outlier analysis) and positioned based on the top BLAST hit to the H. melpomene genome; those that were not confidently or uniquely assigned to these positions are shown as stars (e.g., those at the end of chromosome 10 in D). (Unmapped) Scaffolds of the H. melpomene reference genome that were not assigned to chromosomes in v1.1 of the genome assembly.
Table 2. Accuracy of identification of genomic regions known to control color pattern variation

<table>
<thead>
<tr>
<th>Color pattern loci</th>
<th>B/D</th>
<th>Yb/N/Cr</th>
<th>Ac/Sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome</td>
<td>chr18</td>
<td>chr15</td>
<td>chr10</td>
</tr>
<tr>
<td>Scaffold</td>
<td>HE670865</td>
<td>HE667780</td>
<td>HE668478</td>
</tr>
<tr>
<td>Gene</td>
<td>HMEL001028 (optin)*</td>
<td>Presently unknown</td>
<td>HMEL018000 (WntA)**</td>
</tr>
<tr>
<td>Position</td>
<td>438,423–439,107</td>
<td>450,400–483,854</td>
<td>454,479–454,496</td>
</tr>
<tr>
<td>Functional region^</td>
<td>300,000–400,000</td>
<td>600,000–1,000,000</td>
<td>447,194–454,196</td>
</tr>
<tr>
<td>H. melpomene Peru</td>
<td>assoc</td>
<td>outiler</td>
<td>282,473–376,342</td>
</tr>
<tr>
<td>Position</td>
<td>161,328–376,651</td>
<td>none</td>
<td>376,250</td>
</tr>
<tr>
<td>H. melpomene Ecuador</td>
<td>assoc</td>
<td>outiler</td>
<td>376,651</td>
</tr>
<tr>
<td>Position</td>
<td>263,358</td>
<td>none</td>
<td>367,651</td>
</tr>
<tr>
<td>H. erato Peru</td>
<td>assoc</td>
<td>outiler</td>
<td>376,651</td>
</tr>
<tr>
<td>Position</td>
<td>697,118–725,562</td>
<td>697,118</td>
<td>454,404</td>
</tr>
<tr>
<td>H. erato Ecuador</td>
<td>assoc</td>
<td>outiler</td>
<td>444,914</td>
</tr>
<tr>
<td>Position</td>
<td>362,793–362,794</td>
<td>none</td>
<td>478,952</td>
</tr>
</tbody>
</table>

For each population, positions are given for the SNPs showing the strongest phenotypic associations (assoc) and the highest FST outliers (outlier) on the given scaffold. (N/A) not expected or found; (none) not found.

*From Reed et al. (2011).
**From Martin et al. (2012).
^Inferred from population genomics. The B/D region appears to be similar in H. erato and H. melpomene; Yb/N/Cr region has been localized in H. melpomene only (Nadeau et al. 2012; Supple et al. 2013).

Both the Peruvian and Ecuadorian H. melpomene populations had a SNP at position 97 on an unmapped scaffold, HE670458, that was highly associated with rays (Table 3; Supplemental Table 2). This scaffold appears to consist largely of repetitive elements (BLAST hits match many other regions of the genome), suggesting that there may be a copy of a repetitive element that is associated with the presence of rays in both populations. All rayed individuals were heterozygous, and all nonrayed individuals were homozygous at this SNP in both H. melpomene populations. This would be consistent with multiple alleles aligning to this genomic region or the presence of a unique haplotype sequence linked to the rayed allele. The existence of such a repetitive element is consistent with previous findings that repetitive elements are present in the region of highest divergence at the B/D locus (Papa et al. 2008b; Nadeau et al. 2012).

Surprisingly, in the Peruvian H. erato population the strongest associations with red color pattern elements were not on chromosome 17, but at two scaffolds (HE670771 and HE670235) on chromosome 2 (Fig. 3D; Table 3; Supplemental Fig. 4; Supplemental Table 2). In addition, two SNPs significantly associated with the presence of rays in both populations. All rayed individuals were heterozygous, and all nonrayed individuals were homozygous at this SNP in both H. melpomene populations. This would be consistent with multiple alleles aligning to this genomic region or the presence of a unique haplotype sequence linked to the rayed allele. The existence of such a repetitive element is consistent with previous findings that repetitive elements are present in the region of highest divergence at the B/D locus (Papa et al. 2008b; Nadeau et al. 2012).

Significant associations with yellow in the forewing band were present in the D region in Ecuadorian H. erato (Fig. 4D; Supplemental Table 2), consistent with the fact that the D locus controls both yellow and red coloration in the forewing band in H. erato (Sheppard et al. 1985; Salazar 2012; Papa et al. 2013). No significant associations were found with the presence of yellow color in the forewing band in either Peruvian H. erato or Ecuadorian H. melpomene.
controls the shape as well as the color of the forewing band in Peruvian *H. melpomene*. However, we did also find a cluster of eight SNPs associated with band shape on an unmapped scaffold, HE671554. New mapping analyses suggest that this scaffold is on chromosome 20 (J Davey, pers. comm.) and therefore not linked to any previously described color pattern controlling loci (Table 3).

In Peruvian *H. erato*, the SNP in the *Sd* region that was associated with the yellow hindwing bar also showed the expected association with forewing band shape in the de novo assembly but not the reference alignment. This SNP was just 5 kb upstream of the *WntA* gene (Fig. 3D; Table 2; Supplemental Fig. 4). Associations with forewing band shape were also found on chromosome 2 in this and the Ecuadorian *H. erato* populations (Figs. 3D, 4D; Supplemental Fig. 4), in similar regions to those associated with red color in Peruvian *H. erato* (Table 3; Supplemental Table 2).

In both species from the Ecuadorian hybrid zone, we found SNPs associated with forewing band shape (cell spot 7/8/11) within introns of the *WntA* gene (Fig. 4; Table 2; Supplemental Table 2). In Ecuadorian *H. erato*, we also found two tightly linked SNPs on chromosome 13 and three tightly linked SNPs on an unmapped scaffold (HE669551) that were associated with forewing band shape and also rounding of the band (Fig. 4D; Supplemental Table 2). More recent mapping analysis suggests that both these scaffolds

Figure 4. Association mapping (A,D) and outlier analysis (B,E) for *H. melpomene* (A–C) and *H. erato* (D–F) in Ecuador. See Figure 3 legend for further information.
H. melpomene (0.059%) and H. erato zones showed a greater percentage of SNPs as outliers in both the analyses (Table 1). In the de novo assembly, Peruvian hybrid locality. Reference aligned data from H. melpomene found within each species were more similar than within each assemblies, in the reference alignments the proportions of outliers the reference aligned data was similar. However, unlike the de novo assemblies, in the reference alignments the proportions of outliers found within each species were more similar than within each locality. Reference aligned data from H. melpomene contained ~0.025% outlier SNPs in both Peru and Ecuador, whereas reference aligned data from H. erato had 0.005% outliers in Peru and 0.017% outliers in Ecuador (Table 1). This would be consistent with some of the most rapidly diverging regions being lost in H. erato when aligned against the reference H. melpomene genome.

As suggested by results from the de novo assemblies, there do appear to be differences in population structure between the geographic regions that are consistent across both species. This is also reflected in the FST distributions (from both alignment and assembly approaches), with both H. erato and H. melpomene having higher mean and background levels of FST in Ecuador as compared to Peru (Fig. 2; Table 1; Supplemental Fig. 3), despite the average distance between sampling locations of “pure” subspecies individuals being similar for both hybrid zones (~56 km in Ecuador and 58–60 km in Peru). However, the altitudinal range across the hybrid zone in Ecuador is greater than in Peru (931 m versus 318 m, respectively). Within both regions, H. melpomene has a lower mean FST than H. erato, which would be consistent with higher dispersal distances in H. melpomene, as previously suggested (Mallet et al. 1990). Similar outlier regions were detected by both the alignment and assembly approaches (Figs. 3B,E, 4B,E), although only Peruvian H. melpomene gave a good overlap in the specific SNPs detected (Fig. 5). Some of the outlier contigs detected in Peruvian H. erato could not be positioned on the H. melpomene genome with confidence (Fig. 3B).

Overall, there was considerable overlap between the genomic regions containing outlier SNPs and those showing phenotypic associations (Figs. 3, 4), and to some extent in the specific SNPs, with the majority of phenotypically associated SNPs also being outliers (Fig. 5). The exception to this general trend was the Peruvian H. erato population, where a large proportion of the phenotypically associated SNPs were not strongly divergent between subspecies. In general, the majority of outlier SNPs were within 1 Mb of a known color pattern locus (including the newly identified Ro region; excluding these, 37.5% of outliers in Ecuadorian H. erato were within 1 Mb of the D and Sd loci) (Supplemental Table 2). The strongest outliers on chromosome 10 in the Ecuadorian populations and Peruvian H. erato were within introns of the WntA gene, and the strongest outliers on the B/D scaffold were all 3’ of the optix gene (Table 2; Supplemental Fig. 4).

In both H. melpomene populations there was a second strongly divergent region on chromosome 18 ~2 Mb from the B/D region, which was not divergent in either of the H. erato populations (Fig. 3B; Supplemental Fig. 4). This is the same region on scaffold HE071488 that showed associations with color pattern and altitude in the Peruvian H. melpomene population (Table 3). In the Peruvian H. melpomene population, we detected two clusters of outlier divergent SNPs on chromosome 6, which do not appear to be associated with color pattern (Fig. 3B; Table 3; Supplemental Fig. 4). Outliers were also detected on chromosome 2 in both H. erato populations, some of which were in similar regions to those detected in the association mapping (Table 3; Supplemental Fig. 4).

**Discussion**

It has long been recognized that convergent and parallel evolution provides a natural experimental system in which to study the predictability of adaptation (Stewart et al. 1987; Wood et al. 2005). This approach has come to the fore with the recent integration of molecular and phenotypic studies of adaptive traits (Stinchcombe
and Hoekstra 2007; Nadeau and Jiggins 2010). Here, we have studied parallel divergent clines in two co-mimic species of butterflies, using RAD sequencing to generate an extensive data set covering 1%–5% of the entire genome. Previous genomic studies of these species have sampled only a few individuals of divergent wing races (The species have sampled only a few individuals of divergent wing races; Nadeau et al. 2012), but the location of N and distance from Yb had not been established prior to this study. Moreover, we conducted the first genome-wide scan for divergent loci and identified some that are not wing-color-pattern-related and so may have a role in other aspects of ecological divergence. With these data, we also identify a cryptic population of H. timareta in Ecuador and reveal parallel patterns of divergence between co-mimetic species.

### Comparison of de novo assembly and reference alignment of RAD data

Genome-wide association studies (GWAS) are now common in studies of admixed human populations (Visscher et al. 2012). The use of GWAS studies outside of model organisms has mostly been hampered by lack of reference genomes or methods for typing sufficient numbers of markers. However, these limitations are

---

**Table 3. Novel genomic regions showing phenotypic associations or divergence outliers**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Scaffold</th>
<th>Comparison</th>
<th>Closest gene</th>
<th>Distance</th>
<th>GO function</th>
<th>Putative protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr18</td>
<td>HE671488</td>
<td>melp Peru:</td>
<td>MEL014920</td>
<td>19,171</td>
<td>Catalytic activity, protein binding</td>
<td>Fatty acid synthase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>assoc alt (D gen); outlier melp Ecuador: outlier melp Ecuador:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>assoc rays melp Peru: assoc rays, D gen (alt) erato Peru: assoc D gen (alt, rays, spot 11) erato Ecuador: assoc spot 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmapped</td>
<td>HE670458</td>
<td>melp Ecuador:</td>
<td>No genes on this scaffold, in repetitive region</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chr2</td>
<td>HE670771</td>
<td>erato Peru: assoc D gen (alt, rays, spot 11) erato Ecuador: assoc spot 11</td>
<td>HMEL008318</td>
<td>0 (I)</td>
<td>Catalytic activity, protein binding</td>
<td>Fatty acid synthase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(assoc)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chr2</td>
<td>HE670771</td>
<td>erato Peru: assoc spot 11</td>
<td>HMEL008322</td>
<td>0 (A)</td>
<td>Odorant binding</td>
<td>Odorant binding protein 7</td>
</tr>
<tr>
<td>Chr2</td>
<td>HE670519</td>
<td>erato Peru:</td>
<td>HMEL007059</td>
<td>0 (I/A)</td>
<td>Oxidoreductase activity</td>
<td>3-Dehydroecdysone 3alpha-reductase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>assoc spot 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chr2</td>
<td>HE670235</td>
<td>erato Peru: assoc D gen (alt, rays, spot 11) erato Ecuador: outlier</td>
<td>HMEL005708</td>
<td>56,981</td>
<td>Taste receptor activity</td>
<td>olfactory receptor 4</td>
</tr>
<tr>
<td>Chr2</td>
<td>HE671428</td>
<td>erato Ecuador:</td>
<td>HMEL014154</td>
<td>0 (S)</td>
<td>Choline dehydrogenase activity, oxidoreductase activity, acting on CH-OH group of donors, flavin adenine dinucleotide binding</td>
<td>Glucose dehydrogenase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>outlier</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chr2</td>
<td>HE671428</td>
<td>erato Ecuador:</td>
<td>HMEL014163</td>
<td>0 (A)</td>
<td>Heat shock protein</td>
<td>70</td>
</tr>
<tr>
<td>Chr17</td>
<td>HE671853</td>
<td>erato Ecuador:</td>
<td>HMEL014236</td>
<td>0 (I)</td>
<td>Catalytic activity, serine-type endopeptidase activity</td>
<td>Serine protease 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>assoc HWY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmapped</td>
<td>HE671554</td>
<td>melp Peru: assoc spot 8</td>
<td>HMEL016146</td>
<td>0 (A/S/I)</td>
<td>Protein binding, zinc ion binding</td>
<td>MICAL-like</td>
</tr>
<tr>
<td>(chr20)</td>
<td></td>
<td>erato Ecuador: assoc Ro (spot 7/8); outlier erato Ecuador: assoc spot 11 (Ro, spot 7/8); outlier</td>
<td>HMEL004352</td>
<td>0 (S)</td>
<td>Microtubule binding</td>
<td>Radial spoke head 3</td>
</tr>
<tr>
<td>Unmapped</td>
<td>HE669551</td>
<td>melp Peru:</td>
<td>HMEL009926</td>
<td>3915</td>
<td>Structural constituent of ribosome, RNA binding</td>
<td>Ribosomal protein 54</td>
</tr>
<tr>
<td>(chr13)</td>
<td></td>
<td>erato Ecuador: assoc spot 11 (Ro, spot 7/8); outlier</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chr6</td>
<td>HE671933</td>
<td>melp Peru: outlier</td>
<td>HMEL016074</td>
<td>7925</td>
<td>Oxidoreductase activity</td>
<td>Amine oxidoreductase</td>
</tr>
<tr>
<td>Chr6</td>
<td>HE671934</td>
<td>melp Peru: outlier</td>
<td>HMEL016075</td>
<td>4121</td>
<td>Oxidoreductase activity</td>
<td>Amine oxidoreductase</td>
</tr>
</tbody>
</table>

*a* Analysis in which SNP is detected. (melp) H. melpomene; (erato) H. erato; (outlier) BayeScan FST outlier analysis; (assoc) association analysis with the strongest associated phenotype and additional phenotypes in parentheses; (rays) presence of hindwing rays and fore/hindwing dennis patches; (D gen) predicted B/D genotype; (spot) presence of nonblack color in that wing cell; (alt) altitude; (Ro) rounding of distal edge of forewing band.

*b* If a SNP is within a gene (distance = 0), then the following is noted in parentheses: (A) nonsynonymous; (S) synonymous; (I) within an intron. Further information is given in Supplemental Table 3.
rapidly being eroded as the cost of sequencing decreases and more reference genomes become available. Furthermore, we have shown that alignment of reads to a fairly distantly related reference genome (~15% divergent) can generate meaningful results. In the absence of a reference genome, de novo assembly also detects the same loci, but with somewhat reduced efficacy.

Alignment of sequence reads to the reference genome produced data for more sites, even in the more distantly related species, *H. erato*. One drawback of the Stacks pipeline that we used for de novo assembly of the reads is that it does not assemble and call sequence variants in the paired-end reads. Hence the available sequence for analysis is almost double in the reference alignments as compared to the de novo assembly. However, it also seems that data was lost in the de novo assembly due to divergent alleles not being assembled together. This may have had a larger influence on the *H. erato* assemblies because this species harbors greater genetic diversity than *H. melpomene* (Hines et al. 2011) and so explain why a much lower proportion of the de novo assembled contigs were present across multiple individuals in *H. erato* (Table 1). We also found a higher proportion of variable sites in the reference alignments as compared to the de novo assemblies. This may again be due to poor assembly of the de novo contigs, but it could also represent genetic variability contained in the paired-end reads. It is possible that paired-end reads might be located in more variable regions, particularly if restriction-site-associated reads were biased toward more conserved regions (The *Heliconius* Genome Consortium 2012).

The larger number of SNPs in the reference alignments resulted in larger numbers of outlier and associated SNPs being detected, most of which cluster in the expected genomic regions. Moreover, there appears to be a higher false positive rate in the association mapping using the de novo assembled data. The most likely explanation for this result is that the smaller number of SNPs generated from the de novo assembly gave less power to correct for underlying population structure. Nevertheless, many of the expected associations and outlier regions were detected in the de novo assembled data. The results from assembly and alignment approaches are more concordant in *H. melpomene* than in *H. erato*, particularly at the level of individual SNPs (Fig. 5). This is very likely due to the fact that the *H. melpomene* reference genome was used to generate the sequence alignments in both species. In addition, the lower within-population diversity in *H. melpomene* may also have led to improved de novo assemblies in this species.

Overall, our results suggest that detection of loci underlying adaptive change is likely to be more effective where reads can be mapped to a reference genome. This is perhaps most likely to be the case in populations with high levels of polymorphism, which prevents divergent alleles from assembling. The de novo approach could, and no doubt will, be improved by developing methods that allow paired-end reads to be incorporated into the SNP typing pipeline (Baxter et al. 2011). This would not only allow a higher density of SNPs to be detected but could also improve alignment of divergent alleles. In the meantime, one approach that has been used in other studies is to first perform de novo assembly of RAD-seq reads to generate a consensus reference and then map reads to this reference for SNP calling (Keller et al. 2013).

**Association mapping across hybrid zones is a rapid way of detecting loci underlying phenotypic differences**

We have successfully used association mapping in hybrid zone individuals to identify virtually all the genomic regions known to control color pattern in these populations (Reed et al. 2011, Martin et al. 2012; Nadeau et al. 2012; Supple et al. 2013). It has commonly been supposed that large sample sizes will be necessary in order to identify genes in wild populations. Here we have used recent theoretical predictions from simulated data (Crawford and Nielsen 2013), that for large effect adaptive loci, even small sample sizes can be highly effective in identification of narrow genomic regions underlying adaptive traits (Table 2; Supplemental Fig. 4). We also confirm the prediction that in populations with low background levels of divergence, both divergence outlier and association mapping approaches are effective in detecting regions under divergent selection. In our study, association mapping has the added benefit of identifying the phenotypic effects of the selected loci. One anticipated pitfall of this method was that many of the phenotypes covary across the hybrid zone. However, it appears that with just 10 individuals with admixed phenotypes, we can dissociate most of the variation and thus find distinct genetic associations for known loci. This therefore gives us some confidence that the novel associations that we have detected are real and not due to covariation with other phenotypes.

In Ecuador, we intentionally sampled from sites at the edges of the hybrid zone where both pure and hybrid individuals were present, because we anticipated that individuals from these sites would have the higher levels of admixture between selected alleles. This may explain the clearer patterns observed in *H. erato* in Ecuador as compared to Peru (Figs. 3D, 4D; Supplemental Table 2). In Peruvian *H. erato*, we also find several genomic regions showing phenotypic associations that are not divergence outliers, which may suggest that these are false positives. However, the less-clear signal in Peruvian *H. erato* could also be due to the reduced sample size in this population (27 individuals). Certainly, the reduced number of Ecuadorian *H. melpomene* individuals (22) seems to have reduced the power of the association mapping (Fig. 4B). The lack of any signal at the *Cr* locus in Peruvian *H. erato* is surprising, and may be because the *Cr* associated region is very narrow. There were 789 SNPs present in *H. erato* within the 607-kb region that is associated with the yellow hindwing bar in *H. melpomene*. This is only slightly below the genome-wide average for *H. erato* (~1800 SNPs/Mb), but linkage disequilibrium breaks down rapidly in *H. erato* (Counterman et al. 2010), and so this may not have been sufficient coverage to identify the *Cr* locus. Nevertheless, contrary to previous suggestions (Kromforst et al. 2013), the density of RAD markers we have obtained was sufficient to identify many narrow divergent genomic regions.

Although we have clearly demonstrated the utility of this approach for association mapping, it should be noted that scoring of some phenotypes was informed by previous crossing experiments. For example, the *N* locus in Ecuadorian *H. melpomene* was scored taking into account the genetic background at *B/D* (Salazar 2012), and the scoring of the predicted genotype at the *B/D* locus yielded stronger associations than scoring of individual color pattern elements. Nonetheless, scoring based purely on phenotypic variation did successfully identify color pattern loci in several cases (e.g., *Ro*, *Ac/Sd*, and *Yb*). Overall, the prospects for mapping individual phenotypic components and identifying epistatic relationships without prior knowledge are considerable, especially with larger sample sizes.

The possibility of using hybrid zones for association mapping has long been recognized (Kocher and Sage 1986), but few studies have successfully applied this technique. Studies in younger hybrid zones, for example *Helianthus* sunflowers, have found that linkage disequilibrium between unlinked genomic regions in early generation hybrids can produce spurious associations (Rieseberg...
Heliconius hybrid zones seem ideal in this regard because they appear to be fairly ancient and close to linkage equilibrium. However, it seems likely that many other suitable systems do exist for this type of approach (Lexier et al. 2006; Crawford and Nielsen 2013). An additional benefit of the Heliconius system is that much of the phenotypic variation is controlled by major effect loci, which can be detected with small sample sizes. Although many adaptive phenotypes appear to involve major effect loci (Orr 2005; Nadeau and Jiggins 2010), in order to move beyond these and detect minor effect loci, much larger sample sizes will be required (Beavis 1997). However, by incorporating methods that use a probabilistic framework to infer allele frequencies in low coverage sequencing data (Gompert and Buerkle 2011), it should be feasible to sequence large enough samples for analysis of quantitative traits.

Identification of a novel color pattern locus
Our association mapping results have robustly identified the H. erato Ro locus, that controls the shape of the distal edge of the forewing band, as being on chromosome 13 near gene HE669551. This gene has a predicted Gene Ontology (GO) molecular function of microtubule binding and is similar to other insect Radial Spoke Head 3 proteins, which are components of the cilia (Avidor-Reiss et al. 2004). It is therefore not an obvious candidate for control of color pattern, so may simply be linked to the causative site. Our results are contrary to the suggestion of a recently published QTL study that Ro may be linked to Sd (Papa et al. 2013). However, that study also identified a major unlinked QTL for forewing band shape that could not be assigned to a H. melpomene chromosome and so may be homologous to the locus we detected here. Furthermore, a QTL for several aspects of forewing band shape and size, including the shape of the distal edge, has previously been identified in H. melpomene on chromosome 13 (Baxter et al. 2009). This was located to a fairly broad region but its positioning is consistent with our results for the Ro locus in H. erato. It therefore seems likely that we have identified a new wing patterning locus that is homologous in H. melpomene and H. erato.

Ecological selection across the hybrid zones
Our results support previous assertions that selection acting on color pattern is the most important factor in maintaining these hybrid zones (Mallet and Barton 1989a; Baxter et al. 2010; Counterman et al. 2010; Nadeau et al. 2012; Supple et al. 2013). The most divergent genomic regions correspond to color pattern controlling loci, and at least half of all divergence outliers are in these regions. Nevertheless, some divergent regions do not seem to correspond to color pattern loci and could be candidates for adaptation to other ecological factors. The best candidates appear to be the regions on chromosome 2 in H. erato and chromosome 6 in H. melpomene. The regions on chromosome 2 in the Peruvian H. erato population are also associated with color pattern, but such association could be due to the high covariation of color pattern and sampling location in this population. These regions overlap with predicted genes, including basic metabolic genes and a heat shock protein (Table 3), which could be candidates for adaptation to different temperature regimes. Chemosensory genes were also detected on chromosome 2 and could be candidates for divergent mate preference or host plant adaptation (Briscoe et al. 2013). However, no differences in host plant preference have been observed in Peru, where these outliers were detected, and mating within the hybrid zone appears to be random (Mallet and Barton 1989a), although marginal differences in mate preference have been observed in H. melpomene (Merrill et al. 2011a).

Comparison of the genomic architecture of divergence between convergent species
One interesting question that can be addressed with our results is the extent to which species undergoing parallel divergence will show parallel patterns at the genomic level. In order to address this, we first need to know whether the species really have undergone parallel divergence, i.e., that both the phenotypic start and end points have been similar. Several previous studies have suggested that this is not the case and that H. erato diverged earlier and followed a different trajectory compared to H. melpomene (Brower 1996; Hanagan et al. 2004; Quek et al. 2010). However, our phylogenetic results are more consistent with a recent analysis suggesting that the two species do appear to have undergone codivergence in multiple populations across their range (Cuthill and Charleston 2012). Our results are based on significantly more data than any of the previous analyses (>5 Mb in H. melpomene and >1 Mb in H. erato), and should produce a better signal for phylogenetic analysis as compared to AFLPs used previously (Quek et al. 2010). Although the striking similarities in tree topology do seem to support the codivergence hypothesis, alignment to a reference genome means that the evolutionary rates in our data for H. erato and H. melpomene are not directly comparable. In addition to the phylogenetic signal, our data also suggested similar patterns of population structure between species in each of the regions, with higher background divergence levels in Ecuador as compared to Peru (Fig. 2; Table 1; Supplemental Fig. 3).

Although some loci show parallel divergence in both species (B/D in Peru; B/D and Ac/Sd in Ecuador), there is surprisingly little similarity in the other loci that are divergent when comparing parallel hybrid zones. This is contrary to the general perception that there are strong genetic parallels in this system (Joron et al. 2006; Baxter et al. 2008; Papa et al. 2008a; Supple et al. 2013).
Some of these differences were known previously, for example, that in Peru the Sd/Ac locus controls band shape variation in *H. erato* but not in *H. melpomene* (Mallet 1989). Our results extend this further through the identification of the Ro locus on chromosome 13 in Ecuadorian *H. erato*, which is not divergent in its co-mimic *H. melpomene*, and the identification of divergent regions of chromosome 2 in *H. erato* and chromosome 6 in Peruvian *H. melpomene*.

In general, it seems that although the same color pattern loci are present in both species (Joron et al. 2006; Baxter et al. 2008; Martin et al. 2012), they are being used in different ways and combinations in order to produce convergent phenotypes. This is particularly surprising given the pattern of codivergence observed in the phylogeny, which would appear to suggest that similar color patterns have arisen at a similar time and from similar ancestral forms in both species. Nonetheless, the apparent pattern of codivergence could simply reflect more recent patterns of gene flow between geographically proximate populations in both species. This has recently been highlighted by studies showing that patterns of divergence at color pattern controlling loci can be very different to those found at the rest of the genome (Hines et al. 2011; The Heliconius Genome Consortium 2012; Pardo-Diaz et al. 2012; Supple et al. 2013). Therefore, the differences that we observe in the use of particular loci in the two species could reflect different mimetic histories that will only be resolved by studies of the evolutionary history of particular loci.

**Discovery of a new cryptic *H. timareta* population**

An unexpected finding of our study was the discovery of a previously undescribed population of *H. timareta*, which appears phenotypically virtually indistinguishable from *H. melpomene malletii* in Ecuador but is clearly genetically distinct (Fig. 1; Supplementary Figs. 1, 2). *H. timareta florensis* is a malletii-like population that has previously been described in Colombia and also co-occurs with *H. melpomene malletii*. In that population, the length of the red line on the anterior edge of the ventral forewing was diagnostic (Giraldo et al. 2008). This character was not diagnostic in our genotyped individuals, with overlapping length distributions between the species (data not shown). We noted a tendency toward genotyped individuals, with overlapping length distributions between the species (data not shown). This character was not diagnostic in our study of particular loci in the two species could reflect different mimetic histories that will only be resolved by studies of the evolutionary history of particular loci.

**Conclusions**

We have demonstrated that high-resolution genome scans using admixed individuals from hybrid zones can be used to identify loci underlying phenotypic variation. Only a small proportion of the genome (−0.025%) is strongly differentiated between subspecies, and most of this can be explained by divergence at loci controlling color pattern. This is consistent with previous studies based on smaller numbers of markers (Turner et al. 1979; Baxter et al. 2010; Counterman et al. 2010; Nadeau et al. 2012) and suggests that the hybrid zones are ancient or have formed in primary contact, and are maintained by strong selection on color pattern (Mallet and Barton 1989a; Mallet 2010). However, we also find, for the first time, some divergent loci that do not appear to be associated with color pattern, suggesting that there may be other differences between subspecies. This could explain why several *Heliconius* hybrid zones occur across ecological gradients (Benson 1982), if they are coupled with extrinsic selection acting on other loci in the genome (Bierne et al. 2011). However, this needs to be confirmed with detailed phenotypic analyses of the subspecies to identify whether differences are present that could be explained by ecological adaptation. In general, we find that although some loci are divergent in all populations, the genomic pattern of divergence between mimetic species is not particularly similar, suggesting that the level of parallel genetic evolution between *H. erato* and *H. melpomene* is in fact quite low, despite parallel phylogenetic patterns of divergence. Finally, our analysis shows that alignment to a distantly related reference genome can improve analyses over a de novo assembly of the data.

**Methods**

**Samples and sequencing**

Thirty *H. erato* and 30 *H. melpomene* individuals were selected from a larger sample taken from the hybrid zone region in Peru. Similarly, 30 *H. erato* and 30 *H. melpomene* were also selected from a larger study of a subspecies hybrid zone in Ecuador (Salazar 2012). Each set of 30 samples comprised 10 pure forms of each subspecies and 10 hybrids (based on color pattern). See Figure 2 and Supplemental Table 4 for further details of the samples and locations.

RAD sequencing libraries were prepared using previously described methodologies (Baird et al. 2008; Baxter et al. 2011; The Heliconius Genome Consortium 2012). Briefly, DNA was digested with the restriction enzyme PstI prior to ligation of P1 sequencing adaptors with five-base molecular identifiers (MIDs) (Supplemental Table 4). We then pooled samples into groups of six before shearing, ligation of P2 adaptors, amplification, and fragment size selection (300–600 bp). Libraries were then further pooled such that 30 individuals were sequenced on each lane of an Illumina HiSeq 2000 sequencer to obtain 150 bp paired-end sequences. We obtained an average of 374M sequence pairs from each lane. Following sequencing, three of the *H. erato* individuals from Peru were found to have been incorrectly assigned to this species and were excluded from all further analyses.

In order to compare patterns of phylogenetic divergence of the focal subspecies, we also used sequence data from additional subspecies and closely related species in each group. Two individuals, each from six additional *H. erato* populations and the closely related *H. himera*, were also PstI RAD sequenced with five individuals pooled per lane of Illumina GAIIx (100 bp paired-end
sequencing). These sequences were obtained in the same run as a comparable set of individuals from the *H. melpomene* clade, which have been used in previous analyses and also included *H. cydno, H. timareta,* and *H. hecale* (The Heliconius Genome Consortium 2012; Nadeau et al. 2013) (European Nucleotide Archive, accession ERP000991). We also obtained whole-genome shotgun sequence data from an outgroup species, *H. clysonimus,* which was sequenced on a fifth of a HiSeq 2000 lane, giving 3.5 SM 100 bp read pairs for this individual.

Alignment to reference genome

We separated paired-end reads by MID using the RADpools script in the RADtools (v1.2.4) package (Baxter et al. 2011), which also filters based on the presence of the restriction enzyme cut site, using the option to allow one mismatch within the MID. Reads from each individual were then aligned to the *H. melpomene* reference genome (The Heliconius Genome Consortium 2012) using Stampy v1.0.17 (Lunter and Goodson 2011), with default parameters except substitution rate, which was set to 0.03 for alignments of *H. melpomene* and 0.10 for alignments of *H. erato.*

We then realigned indels and called genotypes using the Genome Analysis Tool Kit (GATK) v1.6.7 (DePristo et al. 2011), outputting all confident sites (those with quality $\geq 30$). This was first run on each set of 30 (or 27) individuals from each population group. These genotype calls were used for analyses of genetic variation within each of the groups, including outlier detection, association mapping, and analyses of subpopulation structure. In addition, genotype calling was also performed on a combined data set of all *H. melpomene* and outgroup taxa (*H. timareta, H. cydno,* and *H. hecale*) as well as a combined set of all *H. erato* and its outgroups (*H. himera* and *H. clysonimus*). These genotype calls were used for the phylogenetic analyses and broader analyses of genetic structure. For all downstream analyses, calls were further filtered to only accept those based on a minimum depth of five reads and minimum genotype and mapping qualities (GQ and MQ) of 30 for *H. melpomene* and 20 for *H. erato.*

De novo assembly

We quality-filtered the single-end raw sequence data and separated sequences by MID with the process_radtags program within Stacks (Catchen et al. 2011). This program corrects single errors in the MID or restriction site and then checks quality score using a sliding window across 15% of the length of the read. We discarded sequences with a raw phred score below 10, removed reads with uncalled bases or low quality scores, and trimmed reads to 100 bases to eliminate potential sequencing error occurring at ends of reads. Table 1 shows the mean read numbers per individual obtained after filtering. For each population group, we assembled loci de novo using the denovo_map.sh pipeline in Stacks (Catchen et al. 2011). We set the minimum depth of coverage (m) to 6, allowed 4 mismatches both in sites with 100% of individuals present for $5_2$ and with a minor allele frequency of at least 20% were retained. This reduced the number of sampled sites, keeping just the most informative ones, for easier handling by the program. Initial short runs (10^3 burn-in, 10^4 data collection, $K = 1$) were used to estimate the allele frequency distribution parameter $\lambda.$ We then ran longer clustering runs (10^4 burn-in, 10^4 data collection) with the obtained values of $\lambda$ for each of the four population groups for $K = 1–3.$ For *H. melpomene* in Ecuador, the analysis was first run with all individuals included and then excluding the individuals identified as being *H. timareta*.

We also performed principal components analysis of the genetic variation in each population group. This was done with the “cmdscale” command in R (R Development Core Team 2011), using genetic distance matrices calculated as 0.5-ibs, where ibs was the identity by sequence matrix calculated in GenABEL (see below). As further confirmation that some of the *H. melpomene* individuals sampled in Ecuador were in fact cryptic *H. timareta,* we also performed principal components analysis on the combined *H. melpomene* and outgroup data set. We also ran principal components analysis on the de novo assembled data for each population group to test whether both methods were detecting similar underlying patterns of genetic variation.

In order to compare our newly identified *H. timareta* individuals to other populations, we Sanger sequenced a 745-bp region of mitochondrial COI that overlapped with the regions sequenced in previous studies (Giraldo et al. 2008; Mérot et al. 2013). This was PCR amplified as in Mérot et al. (2013) with primers “Jerry” and “Patlep” and directly sequenced with “Patlep.” These sequences were then aligned with those available on GenBank, and a maximum likelihood phylogeny was constructed in PhyML (Guindon and Gascuel 2003) with a GTR model and 1000 bootstrap replicates.

Association mapping of loci controlling color pattern variation

We scored components of phenotypic variation that segregate across each of the hybrid zones. The scored phenotypes are shown in Figure 3 (for Peru) and Figure 4 (for Ecuador) and listed in full in Supplementary Table 5. These were scored mostly as binomial (1,0) traits, but in some cases intermediates were also scored (as 0.5). The width and shape of the forewing band was scored based on whether it extended into each of the wing “cells,” demarcated by the major wing veins (as shown in Supplemental Fig. 9). In Peruvian populations, the size and shape of the forewing band was measured as two components (Fig. 3C,F) that extend the band distally (cell spot 8) and proximally (cell spot 11). In Ecuador, three aspects of band shape were scored: cells 8 and 11, which make up the proximal spot in *H. m. plessoni* and *H. e. notabilis,* and cell 7, which pushes the band toward the wing margin in *H. m. malleti* and *H. e. lativitta* (Fig. 4C,F). In our sample of *H. melpomene,* the presence of cell spots 8 and 11 were perfectly correlated, whereas in *H. erato,* the presence of cell spot 7 was perfectly correlated with the absence of cell spot 8. In addition, individuals were also scored for their predicted genotypes at major loci described previously (with

Phylogenetics and analysis of population structure

Only the reference aligned data were used for phylogenetics and Structure analyses. We used custom scripts to convert from VCF to PHYLIP format and to filter sites with a minimum of 95% of individuals with confident calls. Maximum likelihood phylogenies were constructed in PhyML (Guindon and Gascuel 2003) with a GTR model using the resulting 5,737,351 sites (including invariant sites) for the *H. melpomene* group and 1,693,024 sites for the *H. erato* group. Approximate likelihood branch supports were calculated within the program.

Population structure within and across each of the hybrid zones was analyzed using the program Structure v2.3 (Pritchard et al. 2000). We prepared input files using custom scripts, and only sites with 100% of individuals present for *H. melpomene* populations or at least 75% of individuals present for *H. erato* populations and with a minor allele frequency of at least 20% were retained. This reduced the number of sampled sites, keeping just the most informative ones, for easier handling by the program. Initial short runs (10^3 burn-in, 10^4 data collection, $K = 1$) were used to estimate the allele frequency distribution parameter $\lambda.$ We then ran longer clustering runs (10^4 burn-in, 10^4 data collection) with the obtained values of $\lambda$ for each of the four population groups for $K = 1–3.$ For *H. melpomene* in Ecuador, the analysis was first run with all individuals included and then excluding the individuals identified as being *H. timareta.*
predicted heterozygotes scored as 0.5) (Sheppard et al. 1985; Mallet 1989), and the altitude at which they were collected was included as a continuous phenotypic trait.

We performed association mapping using the R Package GenABEL v 1.7-4 (Aulchenko et al. 2007). This was performed on both the de novo assembled and the reference aligned data with a custom script used to convert both from VCF to Illumina SNP format. Individuals identified as being H. timareta were excluded. Filtering was performed within the program to remove sites with >30% missing data and with a minor allele frequency of <3%.

For each population, an analysis of the hindwing ray phenotype using the reference mapped data was first performed using three methods: a straight score test (qtscore), a score test with the first three principal components of genetic variation (calculated as described above) as covariates, and an Eigenstrat analysis (egscore) (Price et al. 2006). The presence of genetic stratification and the ability of these methods to correct for this was analyzed by comparing the inflation factor, \( \lambda \), which is computed by regression in a Q-Q plot to detect genome-wide skew in association values. In all cases, the analyses incorporating population stratification did not give a reduced value of \( \lambda \) and so were not used for subsequent analyses. As our samples were from hybrid zones with >60% of the samples having extreme values of all scored phenotypes, we would expect similar levels of stratification for all phenotypes, so this test for stratification was not repeated for all phenotypes.

We therefore performed score tests for all scored phenotypes across all population groups. Genome-wide significance was determined empirically from 1000 resampling replicates and corrected for population structure using the test specific \( \lambda \) (Supplemental Table S5).

**BayeScan analysis to identify loci under selection**

We used the program BayeScan v2.1 (Foll and Gaggiotti 2008) to look for loci with outlier \( F_S^2 \) values between “pure” individuals of each subspecies type (based on wing color pattern) in each population group. Exclusion of the H. timareta individuals meant that only three pure H. melpomene malleti individuals remained. Therefore, for the purpose of this analysis of H. melpomene in Ecuador, the two hybrid individuals closest to the H. m. malleti side of the hybrid zone (Fig. 2), which also had the most H. m. malleti like phenotypes, were included as H. m. malleti.

The program was run with the prior odds for the neutral model (priors) set to 10, and outlier loci were detected with a false discovery rate (FDR) of 0.05. We ran this analysis using both the de novo assembled and the reference aligned data. Custom scripts were used to convert these to the correct input format. For both analyses, sites were only kept if at least 75% of individuals were sampled for both subspecies in a given comparison.

**Data access**

DNA sequence reads from this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) under accession number ERP003980. COI sequences have been submitted to the EMBL Nucleotide Sequence Database (EMBL-Bank) at ENA under accession numbers HG710096-HG710125. Custom scripts and wing images are available from Data Dryad with doi: 10.5061/dryad.1nc50.

**Acknowledgments**

We thank Simon Baxter, Doug Turnbull, and William Cresko for their help and advice with RAD library preparation and sequencing. Sequencing was performed at the University of Oregon, Genomics Core Facility and The Gene Pool genomics facility in the University of Edinburgh. We would also like to thank Julian Catchen for his help with Stacks. We thank the governments of Peru and Ecuador for their permission to collect and export specimens. Santiago Villamarín from the Museo Ecuatoriano de Ciencias Naturales provided institutional support in Ecuador. We also thank Ismael Aldás, Carlos Robalino, and Patricia Salazar for their assistance with fieldwork. Joanna Riley assisted with DNA extractions. John Davey gave us access to his unpublished mapping results. We thank three anonymous reviewers for their comments. This project was supported by research grants BBSRC H01439X/1, NSF-CREST #0206200, NSF-DEB-1257839, and NSF-IOS 1305686. N.J.N. was funded by a Leverhulme Trust award to C.D.J.; M.R. was funded through the Ford Foundation Postdoctoral Fellowship Program administered by the National Academies; H.O.Z. and J.A.M. were partially supported by NIH-NIGMS INBRE award P20GM103475 and NSF-EPSCoR award 1002410.

**References**


Received November 6, 2013; accepted in revised form May 5, 2014.
Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*


*Genome Res.* 2014 24: 1316-1333 originally published online May 13, 2014
Access the most recent version at doi:10.1101/gr.169292.113