Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing

Citation for published version:

Digital Object Identifier (DOI):
10.1098/rstb.2011.0198

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Philosophical Transactions of the Royal Society B: Biological Sciences

Publisher Rights Statement:
Freely available in PMC.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Research

Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing


1Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK
2CNRS UMR 7205, Muséum National d’Histoire Naturelle, Département Systématique et Évolution, 45 rue Buffon, 75005 Paris, France
3School of Biosciences, University of Exeter in Cornwall, Penryn TR10 9EZ, UK
4Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK
5The Galton Laboratory, University College London, Stephenson Way, London NW1 2HE, UK
6Welcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
7The GenePool Genomics Facility, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK

Heliconius butterflies represent a recent radiation of species, in which wing pattern divergence has been implicated in speciation. Several loci that control wing pattern phenotypes have been mapped and two were identified through sequencing. These same gene regions play a role in adaptation across the whole Heliconius radiation. Previous studies of population genetic patterns at these regions have sequenced small amplicons. Here, we use targeted next-generation sequence capture to survey patterns of divergence across these entire regions in divergent geographical races and species of Heliconius. This technique was successful both within and between species for obtaining high coverage of almost all coding regions and sufficient coverage of non-coding regions to perform population genetic analyses. We find major peaks of elevated population differentiation between races across hybrid zones, which indicate regions under strong divergent selection. These ‘islands’ of divergence appear to be more extensive between closely related species, but there is less clear evidence for such islands between more distantly related species at two further points along the ‘speciation continuum’. We also sequence fosmid clones across these regions in different Heliconius melpomene races. We find no major structural rearrangements but many relatively large (greater than 1 kb) insertion/deletion events (including gain/loss of transposable elements) that are variable between races.

Keywords: Heliconius; colour pattern; divergence; target enrichment; speciation; genomic islands

1. INTRODUCTION

As populations of organisms diverge and eventually become species, regions of the genome under selection will diverge faster than the rest of the genome. The contrast between genomic regions will be enhanced if there is ongoing gene flow between populations, as this will tend to homogenize the background in contrast to regions under divergent selection [1–5]. The current subject of debate is how extensive or important are these ‘genomic islands’ [6–8]. It has been suggested that genomic islands may harbour linked genetic variation which will also diverge between populations because of reduced effective gene flow. This ‘divergence hitchhiking’ could allow other, more weakly selected alleles to accumulate in these regions with reduced between-population recombination, creating blocks of co-selected alleles [9–11]. In this way, these islands might spread as the speciation process continues. However, explicit modelling of these scenarios suggests that early in the divergence process, genomic islands will tend to be small and divergence hitchhiking limited. Once multiple loci are under selection, genome-wide divergence rapidly occurs, facilitated by selection on these loci causing strong reductions in effective gene flow [7]. Furthermore, one of the most widely cited

* Author for correspondence (njn27@cam.ac.uk).


One contribution of 13 to a Theme Issue ‘Patterns and processes of genomic divergence during speciation’.

This journal is © 2011 The Royal Society
empirical examples of ‘genomic islands’ in races of Anopheles mosquitoes has recently been called into question [3]. Thus, both theoretical and empirical support for this view of speciation have recently been subject to scrutiny.

Heliconius butterflies have radiated across the neotropics over about the last 17 Myr, diversifying into over 40 species and many more subspecies and races [12,13]. These multiple levels of divergence with varying levels of gene flow and hybridization make the group ideal for addressing questions of how genomes diverge during speciation [13]. Heliconius butterflies have aposomatic (warning) patterns that indicate their unpalatability to predators. In several species, such as Heliconius melpomene and Heliconius erato, these patterns vary dramatically across their geographical range [14]. In addition, H. melpomene and H. erato are Müllerian mimics and look almost identical wherever they co-occur. This mimicry and the geographical distinctness of the colour pattern races are driven by strong frequency-dependent selection, owing to avoidance by local predators of familiar colour patterns [15,16]. These colour patterns are also commonly involved in mate recognition between populations and species [17–20] and therefore have the potential to drive divergence through increased reproductive isolation between populations. Therefore, the genetic regions controlling colour pattern are under strong divergent selection between populations and so are expected to represent islands of divergence.

Most of the colour pattern variation in these species is controlled by a small number of genetic loci with Mendelian inheritance patterns [14,21]. In addition, in some cases, genes controlling colour pattern and mate preference are co-located [22,23], consistent with the accumulation of genetic differences centred on a few genomic regions. In H. melpomene, small regions on two chromosomes control a disproportionate amount of the colour pattern variation. Linkage group 15 contains three closely linked loci—HmYb, HmSb and HmN—which control most of the yellow and white colour pattern elements [14,24,25]; while linkage group 18 contains two loci—HmB and HmD—controlling most of the red and orange colour pattern elements [14,21,26] (figure 1b shows the colour pattern elements controlled by these loci). Both HmYb/Sb and HmB/D have been mapped to a fine scale and the loci in each group are tightly linked [14,21,24,26]. Convergent colour pattern elements are controlled in H. erato by homologous loci (figure 1c) [26,27]. Similarly, divergent and locally polymorphic ‘tiger’ wing patterns in Heliconius numata are controlled at a single locus, HmP (figure 1), and are associated with two chromosomal inversions, which also map to the same genomic region as HmYb [27,28]. Thus, the same genetic loci regulate wing pattern divergence across Heliconius.

In spite of the availability of sequence and mapping information for these colour pattern loci, comparatively little is known about levels of genetic divergence within and around these loci in wild populations. Previous work has involved sequencing of multiple, sparsely distributed small amplicons from populations across the H. melpomene hybrid zones in Panama and Peru (figure 1); the HmYb and HmB...
genomic regions had significantly elevated differentiation when compared with other regions [29]. However, no nucleotide variation was perfectly associated with colour pattern, suggesting that the functional sites determining colour pattern were not sampled. In addition, levels of differentiation varied stochastically across the regions making it difficult to narrow down regions of interest.

These colour pattern regions are known to regulate both within- and between-species adaptive divergence, and so are prime candidates for studying the unfolding of ‘islands of divergence’ during speciation. Here we characterize variation across these genomic regions more fully both within and between species, using novel techniques to capture, and thus enrich, entire regions of interest prior to high-throughput sequencing [30]. We studied patterns of divergence in the HmYb/Sb and HmB/D regions across a hybrid zone between two races of H. melpomene (H. melpomene aglaope and H. melpomene amaryllis) in Peru and from two further sympatric species (Heliconius timareta and H. numata) at different levels of divergence from H. melpomene (figure 1). Natural hybridization is known among even the most divergent species in this group [13]. Thus, we can test the idea that genomic islands of divergence might widen during and after speciation because of the presence of low levels of continuing gene flow. As far as we are aware, this is the first time such techniques have been used in non-model systems to address ecological questions. One disadvantage of targeted resequencing is that reads are short and so, in general, have to be aligned back to a reference sequence making it difficult to narrow down sites determining colour pattern were not sampled. In addition, levels of differentiation varied stochastically across the regions making it difficult to narrow down regions of interest.

Additional adult H. melpomene individuals were sampled for fosmid library preparation. These comprised one H. m. aglaope and one H. m. amaryllis sampled from either side of the hybrid zone in Peru and single individuals from two races across a hybrid zone in Panama. This hybrid zone has also been extensively studied [29,33] and represents an independent replicate of populations differing at the HmYb locus. Heliconius melpomene rosina is found in Central America and is ‘postman’ patterned, like H. m. amaryllis, while H. m. melpomene is found in north Colombia and lacks a hind-wing yellow bar (figure 1b). One individual of each race was taken from captive populations maintained in the insectaries of the Smithsonian Research Institute in Gamboa, Panama, derived from wild-caught individuals from the Panamá/Colón and Darién provinces of Panama, respectively (figure 1a). Details of sample preservation and DNA extraction methods are given in the electronic supplementary material.

2. MATERIAL AND METHODS

(a) Samples

A hybrid zone between races of H. melpomene in the Department of San Martin, Peru has been studied for many years [15,21]. Four individuals of H. m. aglaope, the rayed lowland form and four H. m. amaryllis individuals, the red and yellow ‘postman’ upland form were sampled from pure populations either side of this narrow hybrid zone (see figure 1 for locations). Single individuals of each of the sympatric species H. n. silvana and H. timareta ssp. nov. were also sampled from within this area. Heliconius numata is a member of the distinct silvaniform clade and therefore is somewhat more distantly related, whereas H. timareta belongs to the Heliconius cydno superspecies, which is either sister to or nested within H. melpomene [12]. The cryptic H. timareta ssp. nov. has only recently been identified in Peru. This species is phenotypically very similar to H. m. amaryllis but can be differentiated on the basis of mitochondrial and some nuclear DNA sequences and consistent morphological differences [31,32].

(b) Target enrichment and sequencing

In order to enrich genomic DNA for regions of interest prior to sequencing, we used the SureSelect system (Agilent Technologies), which uses RNA probes (‘baits’) designed to capture regions of interest from genomic DNA in solution [30]. Our main targets were two genomic regions containing colour pattern switch genes that had previously been sequenced from H. melpomene bacterial artificial chromosome (BAC) clones: one contains the HmYb and HmSb loci and is 1 149 502 bp in length [24], the other is 716 635 bp in length and contains the HmB and HmD loci [26]. We designed 120 base oligo baits with 60 base overlap (two-fold coverage) using OligoTiler (http://tiling.gersteinlab.org/OligoTiler/oligotiler.cgi) to span each of the overlapping BAC clone sequences available for these regions. We also designed baits in a similar way from three further sequenced BACs not linked to these regions (electronic supplementary material, table S1). The BAC clones were derived from a pool of H. melpomene races, so some allelic variation was present between overlapping BACs, although it is not known to which race each BAC sequence belongs. To avoid placing baits in repeated regions, simple repeats, low-complexity regions and H. melpomene-specific repeats described previously [34] were masked using RepeatMasker [35] and a maximum of 10 bp overlap with a repeat region was allowed for any bait. Baits were also designed to span preliminary genomic contig sequences from the ongoing Heliconius genome-sequencing programme (17 230 baits) and randomly selected expressed sequence tag sequences (10 048 baits), although these data will not be discussed here. Following repeat masking, the design directly targeted 3.5 Mb of sequence extracted from 4.5 Mb of genomic reference sequence with 57 610 baits.

Illumina paired-end sequencing libraries with insert sizes of 200–250 bp were prepared for each individual using custom paired-end adaptors incorporating a 5 bp molecular identifying sequence (MID) immediately downstream of the sequencing primer binding site. These were then pooled using equal quantities of DNA in sets of four (additional samples not discussed here.
were also included) prior to being subjected to SureSelect enrichment (Agilent Technologies, SureSelect Target Enrichment System: Illumina Paired-End Sequencing Platform Library Prep, v. 1.0). Each pool was then run on a single lane of an Illumina HiSeq2000 instrument (2.5 pools per lane), and 100 base-paired end data collected. Image analysis and base calling were performed using the ILLUMINA PIPELINE v. 1.7. Reads were sorted by MID, and then trimmed to remove low-quality terminal bases and the MID tags. Reads were trimmed to 94 bases for the first read and 74 for the second as the second read was of lower quality than the first.

(c) **Sequence analysis**

Reads were aligned to the reference sequences of the *HmYb/Sb* and *HmB/D* regions and to the three unlinked BAC sequences using BWA (v. 0.5.8a) [36] with default parameters. Consensus bases and sequence variants were called from the BWA alignments using a Bayesian model implemented in SAMtools (v. 0.1.7) pileup tool [37] in combination with quality filters to exclude read bases with Phred qualities less than 20 (equal to 1% error rate). A low coverage filter was applied to exclude calls from all bases where sequence depth was probably insufficient to provide a high-confidence genotype call. By comparing the number of high-quality single nucleotide polymorphisms (SNPs; here referring to differences from the reference) detected with different coverage level cut-offs, we assessed that a depth of 10 reads per individual was sufficient to detect most high-quality SNPs while removing those of lower quality. As repeat sequences have been incompletely described in *Heliconius*, it was also necessary to apply an upper coverage limit to exclude repetitive regions that were inadvertently captured or sequenced simply because of their high representation in the genome. For population genetics analyses, we removed all positions with more than 200 reads. However, from our assessment of coverage, this is fairly relaxed and did not remove all repeats, and so downstream filters were applied based on the unusually high levels of nucleotide variation in these regions in *H. melpomene* (see below).

Details of analyses of capture efficiency and coverage are given in the electronic supplementary material. The sequence reads and alignments are available from http://main.g2.bx.psu.edu/u/njnadeau/h/heliconius-sure select-june-2011.

(d) **Population genetics analysis**

Alignments between individuals were performed relative to the reference sequence using Galaxy (http://galaxy.psu.edu). Any positions with data missing for any individual in a particular comparison were removed leaving 54.7 per cent of all bases for the colour pattern regions and 31.8 per cent of all bases for the other BAC regions for use in the analysis. Population genetic analyses were performed using custom scripts in R (v. 2.12). Nucleotide diversity (\(\pi\)) was calculated for *H. m. aglaope* and *H. m. amaryllis* at each site as the average proportion of nucleotide differences between all pairs of alleles. Averages across 100 base windows moved by 50 base intervals revealed regions of exceptionally high nucleotide diversity owing to unfiltered repetitive regions. These repeat regions were then removed by filtering out the upper 5 per cent of these 100 base regions based on \(\pi\). This corresponded to a \(\pi\) threshold of 4.3 per cent and was based on manual inspection of a subset of regions found to contain more than two alleles per individual. This left 50.2 per cent of all bases for the colour pattern regions and 31.5 per cent for the other BAC regions. This difference between regions is most probably owing to lower overall bait density of the unlinked BAC regions: overlapping BACs were used to design baits in the colour pattern region contigs, whereas the baits on unlinked BACs were from singletons and not in contigs.

\[F_{ST} = \frac{H_T - H_S}{H_T},\]

where \(H_T\) is the expected heterozygosity in the total population and \(H_S\) is the mean expected heterozygosity of the two races [38]. Expected heterozygosity was calculated based on the Hardy–Weinberg principle as \(2pq\), where \(p\) and \(q\) are allele frequencies among the individuals we sampled of any pair of alleles. Moving averages of \(\pi\) and \(F_{ST}\) were then calculated for 10 kb windows moving in 100 base intervals across the colour pattern regions and the three unlinked BACs. For all moving average/sliding window analyses, windows in which more than 90 per cent of the data were missing (i.e. less than 1 kb were present) were removed.

In order to assess the reproducibility of the \(F_{ST}\) estimates, \(F_{ST}\) was also calculated for every between-race pair of individuals (16 possible pairs). These were compared with the original uncorrected \(F_{ST}\) estimates after subtracting a small sample size correction of 1/\((2S)\) from both the original estimate and the pair-wise estimates, where \(S\) is the sample size in the subpopulation [39]. These values were also compared with all 12 within-race pair-wise \(F_{ST}\) estimates in order to assess the level of divergence owing to within-population sampling error. \(F_{ST}\) was also calculated for the two species-level comparisons: *H. m. aglaope* to *H. timareta* ssp. *nov.* and *H. m. aglaope* to *H. n. silvana*. To make all measures comparable, a sample size correction was again applied with \(S\) calculated as the harmonic mean for the different subpopulation sample sizes (one for *H. timareta* and *H. numata* and four for *H. m. aglaope*).

Background levels of \(F_{ST}\) were estimated from 10 000 bootstrap resampling replicates of 1000 individual nucleotide values (the minimum number of sites with data in each 10 kb window) from the unlinked BACs.

Tajima’s \(D\), a measure of departure from neutrality that can be used to detect selection [40], was also calculated for 10 kb sliding windows across the region as:

\[D = \pi - \theta,\]

where \(\theta\) is the level of nucleotide polymorphism calculated as:

\[\theta = \frac{s}{a},\]
where \( s \) is the number of polymorphic sites divided by
the total number of sites in a given window and where:

\[
a = \frac{n-1}{n} = 2.5929,
\]

where \( n \) is the number of alleles in a sample, in this case
8 for each of the two \( H. melpomene \) races [38]. Nucleotide
divergence was calculated as a measure of divergence for
the three parapatric population/species comparisons:
\( H. m. aglaope \) to \( H. m. amaryllis \), \( H. m. aglaope \) to
\( H. timareta \) ssp. nov. and \( H. m. aglaope \) to \( H. n. silvana \).
Nucleotide divergence was calculated as the mean propor-
tion of nucleotide differences between a given pair of
races or species again for 10 kb sliding windows
across the regions. Mean values of \( \pi \), \( F_{ST} \) Tajima’s \( D \)
and nucleotide divergence for the whole of each of
the colour pattern regions and unlinked BACs were
analyzed with 95% confidence intervals estimated from
1000 bootstrap resampling replicates of individual
nucleotide values.

(e) Fosmid library preparation and sequencing
Sanger sequencing of fosmid clones (of about 35 kb in
size) and de novo assembly was performed from regions
around candidate genes within the colour pattern
regions (see electronic supplementary information for
further details). Fosmid sequences were aligned against
the \( HmYb \) and \( HmB \) BAC ‘walks’ using BLAST
‘Align’ (NCBI) and Artemis Comparison Tool v. 8.0
[41]. The alignments were used to construct a single
\( HmYb \) and \( HmB \) region contig for each race. Pair-wise
sequence alignments of these contigs were made in
CLUSTALW [42].

3. RESULTS

(a) SureSelect targeted resequencing efficiency
and coverage of target regions
Enrichment was successful. We obtained between 9.6
and 24.8 million sequence reads per sample (33 million
paired-end reads per lane). The performance of
\( H. timareta \) was broadly similar to the \( H. melpomene \)
samples and, on average, 33.5% per cent of reads
mapped to the reference sequences. The proportion of
reads which mapped back in the more distantly related
\( H. numata \) was slightly lower at 27.5 per cent (one
sample \( t \)-test \( p = 5.039\times10^{-8} \)). For all samples, more
than 85 per cent of aligned bases mapped to sequence
directly targeted by the baits (for further analysis of
the resequencing efficiency and coverage see the
electronic supplementary material).

(b) Between-race divergence across genomic
‘hotspots’ containing colour pattern genes
The races \( aglaope \) and \( amaryllis \) differ in colour pat-
tterns known to be controlled by \( HmYb \), \( HmN \), \( HmB \)
and \( HmD \) and therefore, we expect to find genetic
differences between them in these regions. Consistent
with previous findings [29], \( F_{ST} \) was significantly elev-
ated (based on bootstrap resampling) within the colour
pattern regions when compared with the unlinked
BAC regions (figure 2 and electronic supplementary
material, table S2). Across both regions, areas of
maximal divergence could be identified and these
appear to extend across regions of about 500 kb in
both cases. In most cases, peaks of maximal divergence
within these regions were found in all individual
pair-wise inter-population comparisons and were not
present in any pair-wise intra-population comparisons.
These peaks are therefore unlikely to be owing to
sampling effects (electronic supplementary material,
figure S1).

Within the \( HmB/D \) region, there are two peaks of
maximal divergence within the sequenced region cor-
responding roughly to genes \( HM00012 \) and \( HM00028 \)
(otix transcription factor [43]; figure 2b). It is possi-
bile that these may represent the two loci \( HmB \) and
\( HmD \). The highest peak of differentiation is near gene
\( HM001012 \), a predicted gene with a product of just 12
amino acids, with no homology to known genes. Also
within this region are one or more repetitive elements,
including one with similarity to a \( Bombyx mori \) non-
long terminal repeat (LTR) retrotransposon. These
repetitive elements are responsible for the missing data
within the peak because there is currently no method
for separating correct unique alignments of reads from
those derived elsewhere in the genome.

Peaks of \( F_{ST} \) are less clear in the \( HmYb \) region with
up to seven peaks, observed with 10 kb moving average
windows, none as high as those observed in the \( HmB/D \)
region (figure 2a,b). One of these corresponds to gene
\( HM00025 \) (fuzzy-like [24]), others lie in the large inter-
genic region between this gene and \( HM00026 \) (parn)
and two correspond to clusters of genes that are outside
of the mapped \( HmYb \) region (which ends before
\( HM00026 \) [24]). These peaks outside the mapped
\( HmYb \) region could be owing to genetic linkage of
non-functional variation or could be owing to the
\( HmN \) locus, which also controls colour pattern variation
across this hybrid zone and is tightly linked to \( HmYb \)
[21], but has not been finely mapped. Using a window
size of 1 kb, peaks of \( F_{ST} \) were as high in \( HmYb \) as in
\( HmB/D \) (electronic supplementary material, figure S2): the
regions of highest divergence appear narrower in
\( HmYb \).

We found only slight reductions in \( \pi \) at sites showing
highest \( F_{ST} \) in the \( HmB/D \) region in both populations,
and only in \( H. m. amaryllis \) in the \( HmYb \) region (elec-
tronic supplementary material, figure S3). However,
there was no reduction in Tajima’s \( D \) when compared
with other genomic regions (electronic supplementary
material, figure S3 and table S2), suggesting that selec-
tion on these regions is sufficiently ancient for levels of
diversity to have been restored by mutation.

(c) Between-species divergence across genomic
‘hotspots’ containing colour pattern genes
We calculated \( F_{ST} \) at two further levels of divergence:
between closely related species \( H. melpomene \) and
\( H. timareta \), which are likely to hybridize relatively fre-
quently [31]; and between the more distantly related
species \( H. melpomene \) and \( H. numata \), which hybridize
very occasionally in the wild [13]. As predicted, the
overall level of divergence both in the colour pattern
regions and in the unlinked BAC regions increases
as along a continuum from race to species (figure 3).
We sequenced only single individuals of *H. timareta* and *H. numata* and so the *F*<sub>ST</sub> values should be interpreted with some caution. However, the peaks between *H. melpomene* races are generally also present between species. Furthermore, most between-species peaks not evident between races are found in both between-species comparisons, suggesting that these are not artefacts of small sample size (figure 3). Between *H. m. aglaope* and *H. timareta*, the major peaks of divergence are in the same positions as those between races of *H. melpomene* and are similar in height. However, levels of divergence between these species remain high outside these regions, with multiple, more disparate, peaks rising above ‘background’ levels as determined from the unlinked BAC regions. This could be evidence for a spreading island of divergence between these species. In contrast, the *H. melpomene* to *H. numata* comparison generally show higher background divergence on the BAC clones (1.2–1.5%) and less difference between colour pattern regions when compared with unlinked regions (0.2–1.0%; table 1).

**Figure 2. Genetic differentiation (*F*<sub>ST</sub>) between *H. m. aglaope* and *H. m. amaryllis* across the colour pattern regions (**HmYb** region (a), **HmB/D** region (b)) and three unlinked BACs (c). *F*<sub>ST</sub> is uncorrected for sample size and calculated as a 10 kb moving average at 100 bp increments. The threshold in the colour pattern regions indicates the upper 95% CI from 10 000 bootstrap resampling replicates of 1000 bp (the minimum number of sites with data in each 10 kb window) of the unlinked BACs. Peaks of *F*<sub>ST</sub> > 0.3 are shaded in grey. Coding regions (black) and introns (grey stripes) are shown at the bottom of the colour pattern regions; annotations of the unlinked BACs were performed using RNAseq data and automated gene prediction (pipeline to be published in the forthcoming genome paper).

**d) Using fosmid sequences to identify insertions/deletions and rearrangements**

Our fosmid sequences were focused on candidate regions highlighted in previous studies [24,29] and so did not cover all of the regions showing high differentiation between races in our targeted sequencing analysis. No large structural rearrangements or inversions were found in any of the comparisons. However, we found multiple regions of sequence misalignment owing to the insertion/deletion of transposable elements.
(up to 5 kb long), as well as some minor rearrangements and sequence duplications (up to 3.7 kb long). Further details are given in the electronic supplementary material, figure S4 and supplementary results.

4. DISCUSSION

(a) Targeted sequence enrichment in a non-model system

We successfully used targeted resequencing to enrich genomic regions containing both coding and non-coding regions. This was performed using probes designed from *H. melpomene* but which also successfully captured sequence from the related species *H. timareta* to *H. m. aglaope* (orange), and between more distantly related species—*H. numata* to *H. m. aglaope* (black). The thresholds are the upper 95% CI from 10,000 bootstrap resampling replicates of 1000 bp of the unlinked BACs. Regions showing peaks of $F_{ST}$ between *H. m. aglaope* and *H. m. amaryllis* are highlighted in pink. Coding regions (dark green) and introns (pale green stripes) are shown.

Figure 3. Genetic differentiation ($F_{ST}$) across the colour pattern regions (a), (b) and three unlinked BACs (c) at three levels of divergence: within-species—*H. m. amaryllis* to *H. m. aglaope* (grey); between closely related species—*H. timareta* to *H. m. aglaope* (orange); and between more distantly related species—*H. numata* to *H. m. aglaope* (black). The thresholds are the upper 95% CI from 10,000 bootstrap resampling replicates of 1000 bp of the unlinked BACs. Regions showing peaks of $F_{ST}$ between *H. m. aglaope* and *H. m. amaryllis* are highlighted in pink. Coding regions (dark green) and introns (pale green stripes) are shown.

(up to 5 kb long), as well as some minor rearrangements and sequence duplications (up to 3.7 kb long). Further details are given in the electronic supplementary material, figure S4 and supplementary results.

4. DISCUSSION

(a) Targeted sequence enrichment in a non-model system

We successfully used targeted resequencing to enrich genomic regions containing both coding and non-coding regions. This was performed using probes designed from *H. melpomene* but which also successfully captured sequence from the related species *H. timareta* and *H. numata*. As far as we are aware, this is the first time that this technique has been used in an evolutionary study across multiple species. There are reasons why targeted resequencing may be expected to be less successful in *Heliconius* than in humans, which these techniques were initially developed for: average GC content is lower in Lepidoptera (35% versus 41% in humans [44,45]) and levels of genetic diversity and indel polymorphism are much higher [29,46], especially since our SureSelect baits were designed from races or species other than those resequenced here. We did find effects of GC content and sequence divergence on sequence coverage of targeted regions. However, we largely overcame these problems by sequencing at high depth. This resulted in high and even coverage of coding regions and sufficient coverage of non-repetitive intergenic regions to perform analyses of selection and divergence. Our successful use of this technique

Phil. Trans. R. Soc. B (2012)
We therefore expect strongly contrasting differences in wing pattern, between populations ining for regions at which there are known to be fixed elements, making this the best candidate for a functional role in pattern specification at the HmB locus [43]. The peaks identified here lie downstream of this gene and upstream of the previously implicated kinesin gene (HM01018). Although the most parsimo-nious hypothesis is that they represent regulatory
regions of the optix gene, we still cannot rule out a functional role for multiple protein-coding genes at HmB/D. It is conspicuous, however, that the major peaks of differentiation found in both regions contain few coding exons, suggesting that regulatory changes are important in both cases. This is consistent with the observation that there are no coding differences between species in the optix gene across Heliconius that have been studied to date [43]. Indeed, the clustering of colour pattern loci in these regions could represent multiple regulatory regions of a single gene in each region. It is tempting to speculate that the transposable element in the centre of the highest divergence peak in HmB/D could be affecting gene regulation; although, we currently have no evidence that it differs between races. The functional targets of selection in the HmYb locus remain elusive, but it is clear that the sequencing approach described here is a powerful method for narrowing down candidate regions. Overall, the data lend support to the argument that changes in regulatory regions are key targets of adaptive evolution [49].

(d) Mimicry loci as islands of divergence
The clustering of multiple loci controlling different colour pattern elements within particular regions is expected to maintain large islands of divergence and could be evidence for divergence hitchhiking in this system. However, the regions of genetic differentiation we find between races are only about 400–600 kb in size, and differentiation appears to drop to background levels beyond this. This is much smaller than inferred regions of differentiation around selected loci in subdivided populations of the pea aphid [9,10], whitefish [50,51] and stickleback [10,52,53]. However, between morphs of Littorina winkles, differentiated regions were only a few kilobases [54], and similar patterns have been inferred from genomic studies of plant speciation, where islands of differentiation are relatively isolated [55]. The explanation put forward for the small extent of the Littorina regions of differentiation was extensive ongoing gene flow, and gene flow between populations during the spread of the selected variant. This is also likely to have been the case in Heliconius, where gene flow across the hybrid zone is high, and the hybrid zones themselves are likely to be mobile [29,56]. This supports theoretical predictions that when gene flow is high, divergently hitchhiking regions tend to be small and new beneficial mutations are unlikely to be captured [7,11].

On the other hand, between the sympatric species, Heliconius melpomene and H. timareta, we found more peaks of divergence over a larger genomic area. This type of pattern has been suggested to indicate divergence hitchhiking [9,10]. These species are separated by strong pre-mating isolation and in other populations it has been shown that mate preferences are genetically associated with wing-patterning genes [23]. This is consistent with the hypothesis that loci affecting mate choice are more likely to diverge if located within the divergent regions caused by selection acting on the colour pattern genes [9–11]; although if the selective coefficient favouring the new preference mutation is of the order or higher than the effective migration rate, these associations are likely to be fortuitous and not primarily driven by divergence hitchhiking [11]. Overall, our results suggest that the region of influence of wing-patterning loci broadens progressively between hybridizing species, rather than contributing greatly to genome-wide increase in reproductive isolation, although broader sampling would be necessary to confirm this pattern.

In contrast, there was less striking evidence for islands of divergence between H. melpomene and the more distantly related H. numata, but rather a generally higher level of background divergence. Therefore, these regions may have been less important in the divergence of these species, or increased isolation throughout the genome could have started to obscure the 'islands of divergence' pattern. Overall, the data support theoretical predications that divergence hitchhiking is most likely to operate at intermediate stages of speciation [11]. Our findings are also similar to those in lake whitefish, where the size of divergent regions increases with increasing reproductive isolation [51]. Clearly, further sampling of more Heliconius taxa across the ‘speciation continuum’ and more individuals per taxon will be necessary to verify these findings. In addition, sampling of more loci throughout the genome will be necessary to improve estimates of background divergence and to perform full outlier analyses.

5. CONCLUSIONS AND FUTURE DIRECTIONS
The novel techniques applied here provide the most complete picture to date of how selection generates divergence at a genomic scale between hybridizing taxa in this system. The peaks of divergence we observe greatly narrow the candidate regions under divergent selection, paving the way for understanding functional variation. As sequencing costs continue to plummet, sequencing more individuals and races will allow us to narrow down these regions further. We have also identified clearly demarcated islands of divergence among races and species, implying that gene flow often homogenizes regions outside. Sequencing of the Heliconius melpomene genome is currently underway, as are several RAD sequencing projects (similar to those described elsewhere is this issue [53]) and whole-genome resequencing of multiple races and species. These studies will reveal the extent to which colour pattern regions are divergence outliers in the context of the whole genome.

The targeted resequencing data used in this paper are deposited in the European Nucleotide Archive under study ERP000971. Fosmid sequences are deposited in GenBank with accessions: (i) FP700056, FP578989, FP700120, FP884227; (ii) FP700121, FP884228; (iii) FP578990; (iv) FP700117, FP884224; (v) FP885842, FP885850; (vi) FP565936, FP924937; (vii) FP885843; (viii) FP700119; (ix) FP700055; (x) FP884222; (xi) FP885849, FP885844; (xii) FP884221, FP700057; (xiii) FP700053; (xiv) FP565804; (xv) FP700104, FP700054; (xvi) FP884226, FP884225; (xvii) FP884223 (numerals refer to contigs as indicated in figure S4).

This work was funded primarily by a Leverhulme Trust award to C.D.J. and a BBSRC grant to J.M., C.D.J. and M.B. Funding for A.W. was by the European Research
Council starting grant ‘MimEvol’ to M.J. and R.T.J. was funded by a Leverhulme grant to R.H.F.-C. The GenePool is funded by awards from the MRC, NERC, the Scottish Universities’ Life Sciences Alliance and the Darwin Trust of Edinburgh. We thank the Dirección General Forestal y de Fauna Silvestre (Ministerio de Agricultura) in Peru and ANAM in Panama for collecting permits and the staff at the Smithsonian Tropical Research Institute insectaries in Gamboa Panama. We also thank Alexi Balmuth for performing the SureSelect target enrichment and the sequencing teams at the GenePool, Edinburgh (SureSelect sequencing) and the Wellcome Trust Sanger Institute (preparing sequencing founds). Thanks are also due to Paul Wilkinson for help with repeat masking, Jamie Walters for annotating additional BACs and Vincent Plagnol for helpful discussions and informatics support.

REFERENCES


next-generation sequencing. Nat. Meth. 7, 111–118. (doi:10.1038/nmeth.1419)