PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2012) **367**, 343–353
doi:10.1098/rstb.2011.0198

Research

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

Nicola J. Nadeau[1,*], Annabel Whibley[2], Robert T. Jones[2,3], John W. Davey[4], Kanchon K. Dasmahapatra[5], Simon W. Baxter[1], Michael A. Quail[6], Mathieu Joron[2], Richard H. ffrench-Constant[3], Mark L. Blaxter[4,7], James Mallet[5] and Chris D. Jiggins[1]

[1]*Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK*
[2]*CNRS UMR 7205, Muséum National d'Histoire Naturelle, Département Systématique et Evolution, 45 rue Buffon, 75005 Paris, France*
[3]*School of Biosciences, University of Exeter in Cornwall, Penryn TR10 9EZ, UK*
[4]*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK*
[5]*The Galton Laboratory, University College London, Stephenson Way, London NW1 2HE, UK*
[6]*Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK*
[7]*The 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).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2011.0198 or via http://rstb.royalsocietypublishing.org.

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

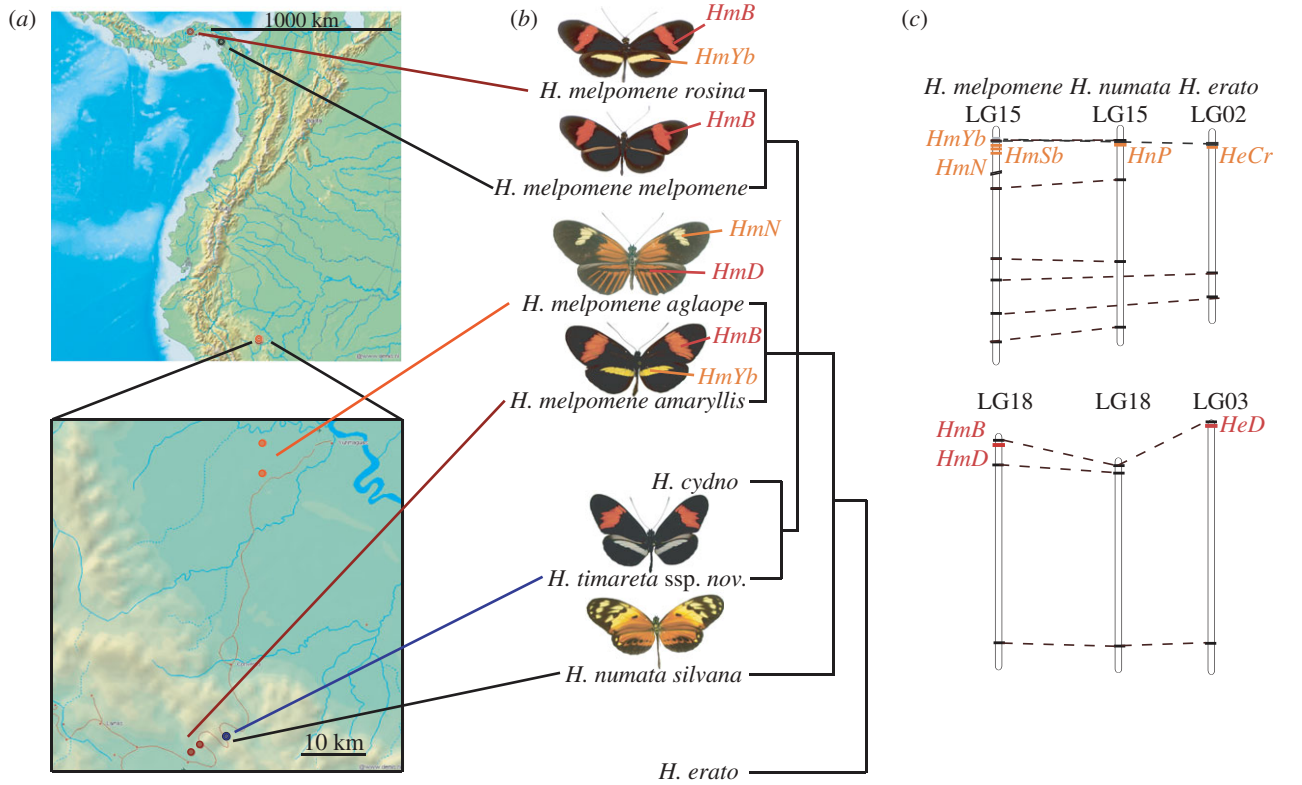

Figure 1. Sampling sites, phylogenetic relationships and colour pattern loci. (*a*) Sampling sites of butterflies used for creating fosmid libraries and for target enrichment sequencing. (*b*) The phylogenetic relationships of the *Heliconius* species referred to in the text. (*c*) Chromosome maps showing the relative positions of the colour pattern loci that are referred to (modified from Baxter *et al.* [26] and Joron *et al.* [27]). The wing colour pattern elements controlled by the *H. melpomene* loci are also indicated in (*b*).

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 aposematic (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 1*b* 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 1*c*) [26,27]. Similarly, divergent and locally polymorphic 'tiger' wing patterns in *Heliconius numata* are controlled at a single locus, *HnP* (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 hard to identify variation in transposable elements, large insertions/deletions and genome rearrangements. Therefore, we also created fosmid libraries for four races of *H. melpomene* and sequenced clones from these to create high-quality race-specific sequences spanning large portions of the colour pattern regions centred on candidate gene loci, to survey events associated with colour pattern differences.

2. MATERIAL AND METHODS

(a) *Samples*

A hybrid zone between races of *H. melpomene* in the Department of San Martín, 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].

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 1*b*). 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 1*a*). Details of sample preservation and DNA extraction methods are given in the electronic supplementary material.

(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 artifical 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 (π) 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 π. This corresponded to a π 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}, a statistic used to measure genetic differentiation between sub-populations, was calculated between *H. m. aglaope* and *H. m. amaryllis* for each nucleotide position using the equation:

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

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 π 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 θ 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:

$$a = \sum_{i=1}^{n-1} \frac{1}{i} = 2.5929,$$

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 proportion of nucleotide differences between a given pair of races or species again for 10 kb sliding windows across the regions. Mean values of π, F_{ST}, Tajima's D and nucleotide divergence for the whole of each of the colour pattern regions and unlinked BACs were calculated 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 $p = 5.039e^{-08}$). 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 patterns 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 elevated (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 corresponding roughly to genes *HM01012* and *HM01028* (*optix* transcription factor [43]; figure 2*b*). It is possible that these may represent the two loci *HmB* and *HmD*. The highest peak of differentiation is near gene *HM01012*, 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 2*a*,*b*). One of these corresponds to gene *HM00025* (*fizzy-like* [24]), others lie in the large intergenic 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 π at sites showing highest F_{ST} in the *HmB/D* region in both populations, and only in *H. m. amaryllis* in the *HmYb* region (electronic 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 selection 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 frequently [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).

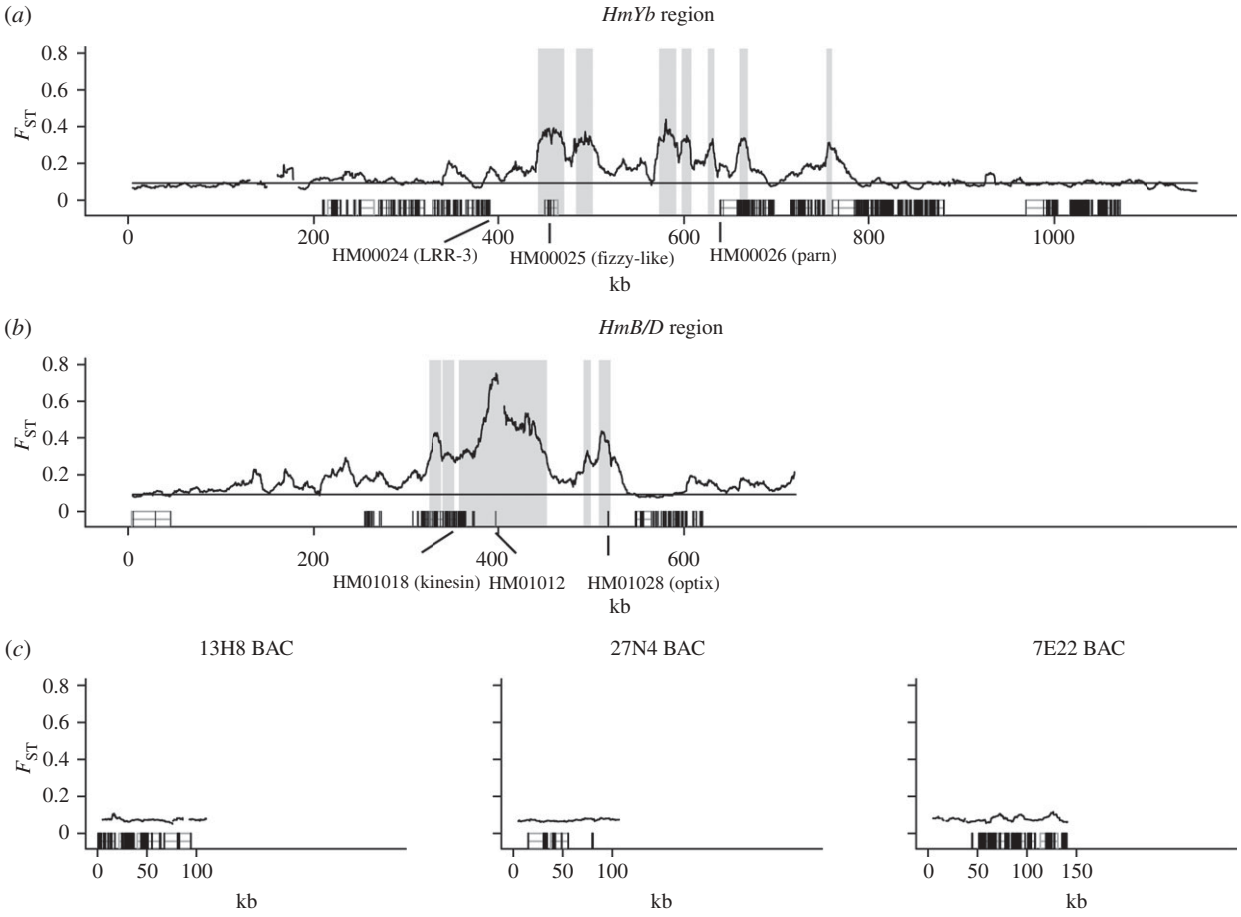

Figure 2. Genetic differentiation (F_{ST}) between *H. m. aglaope* and *H. m. amaryllis* across the colour pattern regions (*HmYb* region (*a*), *HmB/D* region (*b*)) and three unlinked BACs (*c*). F_{ST} is uncorrected for sample size and calculated as a 10 kb moving average at 100 bp increments. The 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 $F_{ST} > 0.3$ 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).

We sequenced only single individuals of *H. timareta* and *H. numata* and so the F_{ST} 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 has far fewer peaks that rise above 'background' levels and these are generally lower and less extensive than those seen in the other comparisons. Overall, the peaks of divergence found between races of *H. melpomene* and between *H. melpomene* and *H. timareta* are as high as those found between *H. melpomene* and *H. numata*, showing that within these regions, gene flow is reduced to the level of reproductively isolated species even between taxa with high background levels of gene flow.

Similar patterns are found when looking at levels of nucleotide divergence. Between *H. m. aglaope* and *H. timareta*, average nucleotide divergence across the colour pattern regions is significantly higher than in unlinked BAC regions (by 0.4–1.2%; table 1), and is also higher than it is in the within-species comparison (by 0.3–0.5%). In contrast, *H. melpomene* to *H. numata* comparisons 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).

(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

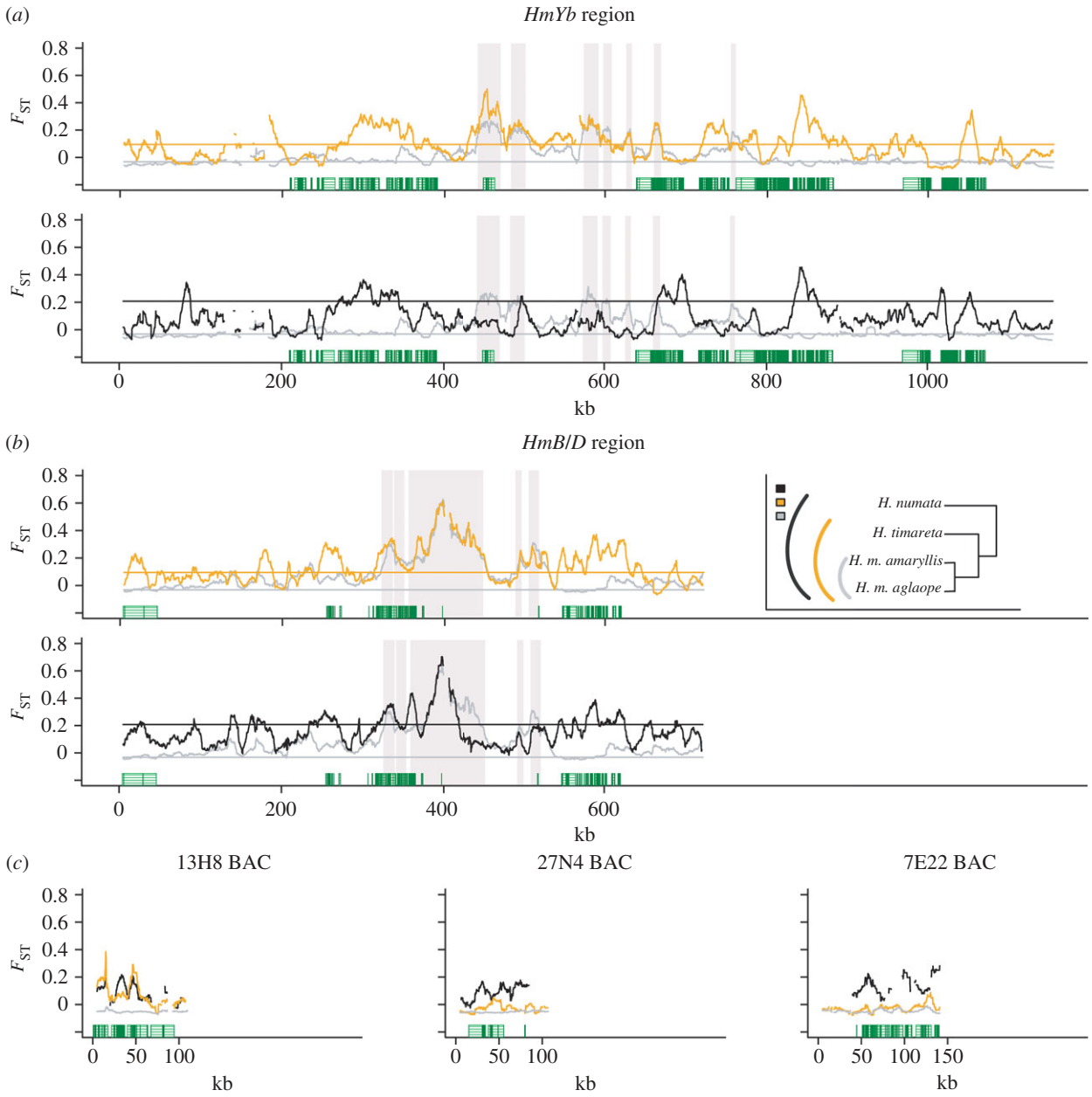

Figure 3. Genetic differentiation (F_{ST}) across the colour pattern regions (*a*,*b*) and three unlinked BACs (*c*) at three levels of divergence: within-species—*H. m. amaryllis* to *H. m. aglaope* (grey); between closely related species—*H. timareta* to *H. m. aglaope* (orange); and between more distantly 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. agaope* 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

Table 1. Per cent nucleotide divergence across the colour pattern regions (*HmYb* and *HmB/D*) and across the three unlinked BAC sequences (hm13H8, hm27N4 and hm7E22). Values are means for the whole of each of the regions ± 95% CI calculated from 1000 bootstrap resampling replicates of individual nucleotide values.

	HmYb	*HmB/D*	hm13H8	hm27N4	hm7E22
H. m. aglaope/*H. m. amaryllis*	1.37 (±0.02)	1.58 (±0.02)	0.96 (±0.05)	0.91 (±0.05)	0.91 (±0.06)
H. m. aglaope/*H. timareta* ssp. *nov.*	1.82 (±0.03)	1.99 (±0.04)	1.25 (±0.10)	0.97 (±0.07)	0.91 (±0.06)
H. m. aglaope/*H. numata*	1.86 (±0.03)	2.03 (±0.04)	1.50 (±0.14)	1.22 (±0.10)	1.39 (±0.13)

therefore demonstrates great potential for application in other non-model systems of evolutionary and ecological interest.

(b) *Characterization of selection and divergence at mimicry loci*

Using a 10 kb sliding window analysis, we identified one or possibly two clear peaks of divergence in the *HmB/D* region and a region of generally elevated divergence in the *HmYb* region. This represents a significant advance on previous attempts to characterize genes responsible for colour pattern divergence in these regions [24,29]. Peaks of differentiation identified in previous work lie on the shoulder of the broader peaks identified here [29]. The regions identified here were not sampled previously primarily because they lie in non-coding regions and were difficult to amplify. Another reason why apparent peaks identified previously were misleading is that on a fine scale, F_{ST} is highly variable. For example, when a 1 kb window is used to calculate F_{ST} (electronic supplementary material, figure S2), there is a great deal of stochasticity even across the regions of peak divergence, indicating that sampling 1 kb amplicons by traditional Sanger sequencing can lead to misleading results. Although this stochasticity may be accentuated by small sample size at these small spatial scales, it was also evident in the previous study particularly in the *HmYb* region. It is also predicted that F_{ST} will not show a smooth linear incline towards a point of highest F_{ST} [47] owing to variation in the level of ancestral polymorphism and differences in the level of background selection [9,10].

Overall, we believe that the current approach based on more complete sampling of the genomic region is more powerful despite the small sample sizes of individuals. Small sample sizes lead to a potential for high background noise in F_{ST} estimates. However, averaging across 10 kb windows reduces much of the noise produced by sampling effects. These windows are on a larger scale than the decline of linkage disequilibrium, which occurs over about 1 kb, such that large windows will average across many independently evolving regions. Consistent with this, we find that peaks of divergence are repeatedly identified when comparing all pairs of individuals between populations. This implies that high divergence in these same regions will be identified when larger samples are investigated. It should be noted that we are searching for regions at which there are known to be fixed differences in wing pattern, between populations with extensive gene flow in the rest of the genome. We therefore expect strongly contrasting F_{ST} values between adjacent genomic regions, and this is indeed what is observed. Small sample sizes such as this may not be adequate to detect more subtle outlier effects in other systems.

Nonetheless, small sample size will lead to F_{ST}, if interpreted as an estimator of a ratio of variances, being overestimated by approximately $1/(2S)$ under the null hypothesis of no differentiation, where S is the subpopulation sample size [39]. This will not affect the use of F_{ST} as a statistic to identify regions of highest divergence, even though our values of F_{ST} may not be comparable with those in other studies using sample-size corrections. Our measurements of F_{ST} across the entire colour pattern regions are lower than the means calculated from small candidate amplicons by Baxter *et al.* [29] (0.132 versus 0.158 for *HmYb* and 0.184 versus 0.281 for *HmB/D*), while our values for individual regions are similar, suggesting that the estimates of F_{ST} in these regions are comparable.

Our results show differences between colour pattern loci, consistent with previous analyses [29], with *HmB/D* region differentiation generally stronger and more consistent over a larger region when compared with *HmYb*. This could be owing to differences in recombination rate, which may be up to twofold lower in *HmB/D* when compared with *HmYb* [24,26]. However, an alternative is stronger selection at *HmB/D*, perhaps because the alleles at these two linked loci show a co-dominant pattern of inheritance, with hybrids expressing both red elements. In contrast, dominant alleles at the linked loci, *HmYb* and *HmN*, are inherited together (i.e. both the *HmYb* allele for absence of a yellow hindwing bar and the *HmN* allele for presence of a forewing yellow band are dominant and found together in *H. m. aglaope*), suggesting that selection on these loci could be weaker, particularly in one direction across the hybrid zone. Nonetheless, the slopes of morphological single-locus clines in the hybrid zone are similar between loci, suggesting a similar strength of selection across loci [48].

(c) *Candidate genes for functional control of wing pattern*

Our results are concordant with the recent discovery of striking patterns of expression of the *optix* transcription factor (*HM01028*) in association with red wing pattern 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 parsimonious 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 *H. melpomene*, 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, *H. 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 *H. 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.Hf.-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 and sequencing fosmids). 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

1 Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. 2009 Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18**, 375–402. (doi:10.1111/j.1365-294X.2008.03946.x)

2 Emelianov, I., Marec, F. & Mallet, J. 2004 Genomic evidence for divergence with gene flow in host races of the larch budmoth. *Proc. R. Soc. Lond. B* **271**, 97–105. (doi:10.1098/rspb.2003.2574)

3 Lawniczak, M. K. N. *et al.* 2010 Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* **330**, 512–514. (doi:10.1126/science.1195755)

4 Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. 2005 Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**, e285. (doi:10.1371/journal.pbio.0030285)

5 Wu, C.-I. 2001 The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865. (doi:10.1046/j.1420-9101.2001.00335.x)

6 Feder, J. L. & Nosil, P. 2009 Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? *Evolution* **63**, 3061–3075. (doi:10.1111/j.1558-5646.2009.00786.x)

7 Feder, J. L. & Nosil, P. 2010 The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **64**, 1729–1747. (doi:10.1111/j.1558-5646.2009.00943.x)

8 Nosil, P. & Feder, J. L. 2012 Genomic divergence during speciation: causes and consequences. *Phil. Trans. R. Soc. B* **367**, 332–342. (doi:10.1098/rstb.2011.0263)

9 Via, S. & West, J. 2008 The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.* **17**, 4334–4345. (doi:10.1111/j.1365-294X.2008.03921.x)

10 Via, S. 2012 Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Phil. Trans. R. Soc. B* **367**, 451–460. (doi:10.1098/rstb.2011.0260)

11 Feder, J. L., Gejji, R., Yeaman, S. & Nosil, P. 2012 Establishment of new mutations under divergence and genome hitchhiking. *Phil. Trans. R. Soc. B* **367**, 461–474. (doi:10.1098/rstb.2011.0256)

12 Beltran, M., Jiggins, C. D., Brower, A. V., Bermingham, E. & Mallet, J. 2007 Do pollen feeding, pupal-mating and larval gregariousness have a single origin in *Heliconius* butterflies? Inferences from multilocus DNA sequence data. *Biol. J. Linn. Soc.* **92**, 221–239. (doi:10.1111/j.1095-8312.2007.00830.x)

13 Mallet, J., Beltran, M., Neukirchen, W. & Linares, M. 2007 Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evol. Biol.* **7**, 28. (doi:10.1186/1471-2148-7-28)

14 Sheppard, P. M., Turner, J. R. G., Brown, K. S., Benson, W. W. & Singer, M. C. 1985 Genetics and the evolution of Muellerian mimicry in *Heliconius* butterflies. *Phil. Trans. R. Soc. Lond. B* **308**, 433–610. (doi:10.1098/rstb.1985.0066)

15 Mallet, J. & Barton, N. H. 1989 Strong natural selection in a warning-color hybrid zone. *Evolution* **43**, 421–431. (doi:10.2307/2409217)

16 Kapan, D. D. 2001 Three-butterfly system provides a field test of Mullerian mimicry. *Nature* **409**, 338–340. (doi:10.1038/35053066)

17 Jiggins, C. D., Naisbit, R. E., Coe, R. L. & Mallet, J. 2001 Reproductive isolation caused by colour pattern mimicry. *Nature* **411**, 302–305. (doi:10.1038/35077075)

18 Jiggins, C. D., Estrada, C. & Rodrigues, A. 2004 Mimicry and the evolution of premating isolation in *Heliconius melpomene* Linnaeus. *J. Evol. Biol.* **17**, 680–691. (doi:10.1111/j.1420-9101.2004.00675.x)

19 Mavárez, J., Salazar, C. A., Bermingham, E., Salcedo, C., Jiggins, C. D. & Linares, M. 2006 Speciation by hybridization in *Heliconius* butterflies. *Nature* **441**, 868–871. (doi:10.1038/nature04738)

20 Merrill, R. M., Gompert, Z., Dembeck, L. M., Kronforst, M. R., McMillan, W. O. & Jiggins, C. D. 2011 Mate preference across the speciation continuum in a clade of mimetic butterflies. *Evolution* **65**, 1489–1500. (doi:10.1111/j.1558-5646.2010.01216.x)

21 Mallet, J. 1989 The genetics of warning colour in Peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proc. R. Soc. Lond. B* **236**, 163–185. (doi:10.1098/rspb.1989.0019)

22 Kronforst, M. R., Young, L. G., Kapan, D. D., McNeely, C., O'Neill, R. J. & Gilbert, L. E. 2006 Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proc. Natl Acad. Sci. USA* **103**, 6575–6580. (doi:10.1073/pnas.0509685103)

23 Merrill, R. M., Van Schooten, B., Scott, J. A. & Jiggins, C. D. 2011 Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. *Proc. R. Soc. B* **278**, 511–518. (doi:10.1098/rspb.2010.1493)

24 Ferguson, L. *et al.* 2010 Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Mol. Ecol.* **19**, 240–254. (doi:10.1111/j.1365-294X.2009.04475.x)

25 Jiggins, C. D., Mavarez, J., Beltrán, M., McMillan, W. O., Johnston, J. S. & Bermingham, E. 2005 A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics* **171**, 557–570. (doi:10.1534/genetics.104.034686)

26 Baxter, S. W. *et al.* 2008 Convergent evolution in the genetic basis of Mullerian mimicry in *Heliconius* butterflies. *Genetics* **180**, 1567–1577. (doi:10.1534/genetics.107.082982)

27 Joron, M., Papa, R., Beltrán, M., Chamberlain, N., Mavárez, J., Baxter, S., ffrench-Constant, R. H., McMillan, W. O. & Jiggins, C. D. 2006 A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol.* **4**, e303. (doi:10.1371/journal.pbio.0040303)

28 Joron, M. *et al.* 2011 Chromosomal rearrangements maintain single-locus polymorphic mimicry. *Nature* **477**, 203–206. (doi:10.1038/nature10341)

29 Baxter, S. W. *et al.* 2010 Genomic hotspots for adaptation: the population genetics of Mullerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet.* **6**, e1000794. (doi:10.1371/journal.pgen.1000794)

30 Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. & Turner, D. J. 2010 Target-enrichment strategies for

next-generation sequencing. *Nat. Meth.* **7**, 111–118. (doi:10.1038/nmeth.1419)
31 Mallet, J. 2009 Rapid speciation, hybridization and adaptive radiation in the *Heliconius melpomene* group. In *Speciation and patterns of diversity* (eds R. Butlin, J. Bridle & D. Schluter), pp. 177–194. Cambridge, UK: Cambridge University Press.
32 Jiggins, C. D. 2008 Ecological speciation in mimetic butterflies. *BioScience* **58**, 541–548. (doi:10.1641/B580610)
33 Mallet, J. 1986 Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity* **56**, 191–202. (doi:10.1038/hdy.1986.31)
34 Otto, T. D., Gomes, L. H. F., Alves-Ferreira, M., de Miranda, A. B. & Degrave, W. M. 2008 ReRep: computational detection of repetitive sequences in genome survey sequences (GSS). *BMC Bioinform.* **9**, 366. (doi:10.1186/1471-2105-9-366)
35 Smit, A. F. A., Hubley, R. & Green, P. 1996 RepeatMasker. See http://www.repeatmasker.org.
36 Li, H. & Durbin, R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)
37 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
38 Hartl, D. L. & Clark, A. G. 1997 *Principles of population genetics*, 3rd edn. Sunderland, MA: Sinauer Associates.
39 Waples, R. 1998 Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *J. Hered.* **89**, 438–450. (doi:10.1093/jhered/89.5.438)
40 Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
41 Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G. & Parkhill, J. 2005 ACT: the Artemis Comparison Tool. *Bioinformatics* **21**, 3422–3423. (doi:10.1093/bioinformatics/bti553)
42 Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. & Thompson, J. D. 2003 Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–3500. (doi:10.1093/nar/gkg500)
43 Reed, R. D. *et al.* 2011 *Optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**, 1137–1141. (doi:10.1126/science.1208227)
44 Lander, E. S. *et al.* 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. (doi:10.1038/35057062)
45 d' Alençon, E. *et al.* 2011 Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proc. Natl Acad. Sci. USA* **107**, 7680–7685. (doi:10.1073/pnas.0910413107)
46 Li, W.-H. & Sadler, L. A. 1991 Low nucleotide diversity in man. *Genetics* **129**, 513–523.
47 Charlesworth, B., Nordborg, M. & Charlesworth, D. 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70**, 155–174. (doi:10.1017/S0016672397002954)
48 Mallet, J., Barton, N., Lamas, G., Santisteban, J., Muedas, M. & Eeley, H. 1990 Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics* **124**, 921–936.
49 Stern, D. L. 2000 Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079–1091.
50 Rogers, S. & Bernatchez, L. 2007 The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonus sp.* Salmonidae) species pairs. *Mol. Biol. Evol.* **24**, 1423–1438. (doi:10.1093/molbev/msm066)
51 Renaut, S., Maillet, N., Normandeau, E., Sauvage, C., Derome, N., Rogers, S. M. & Bernatchez, L. 2012 Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Phil. Trans. R. Soc. B* **367**, 354–363. (doi:10.1098/rstb.2011.0197)
52 Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A. & Cresko, W. A. 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862. (doi:10.1371/journal.pgen.1000862)
53 Hohenlohe, P. A., Bassham, S., Currey, M. & Cresko, W. A. 2012 Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Phil. Trans. R. Soc. B* **367**, 395–408. (doi:10.1098/rstb.2011.0245)
54 Wood, H. M., Grahame, J. W., Humphray, S., Rogers, J. & Butlin, R. K. 2008 Sequence differentiation in regions identified by a genome scan for local adaptation. *Mol. Ecol.* **17**, 3123–3135. (doi:10.1111/j.1365-294X.2008.03755.x)
55 Strasburg, J. L., Sherman, N. A., Wright, K. M., Moyle, L. C., Willis, J. H. & Rieseberg, L. H. 2012 What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Phil. Trans. R. Soc. B* **367**, 364–373. (doi:10.1098/rstb.2011.0199)
56 Mallet, J. 2010 Shift happens! Shifting balance and the evolution of diversity in warning colour and mimicry. *Ecol. Entomol.* **35**, 90–104. (doi:10.1111/j.1365-2311.2009.01137.x)