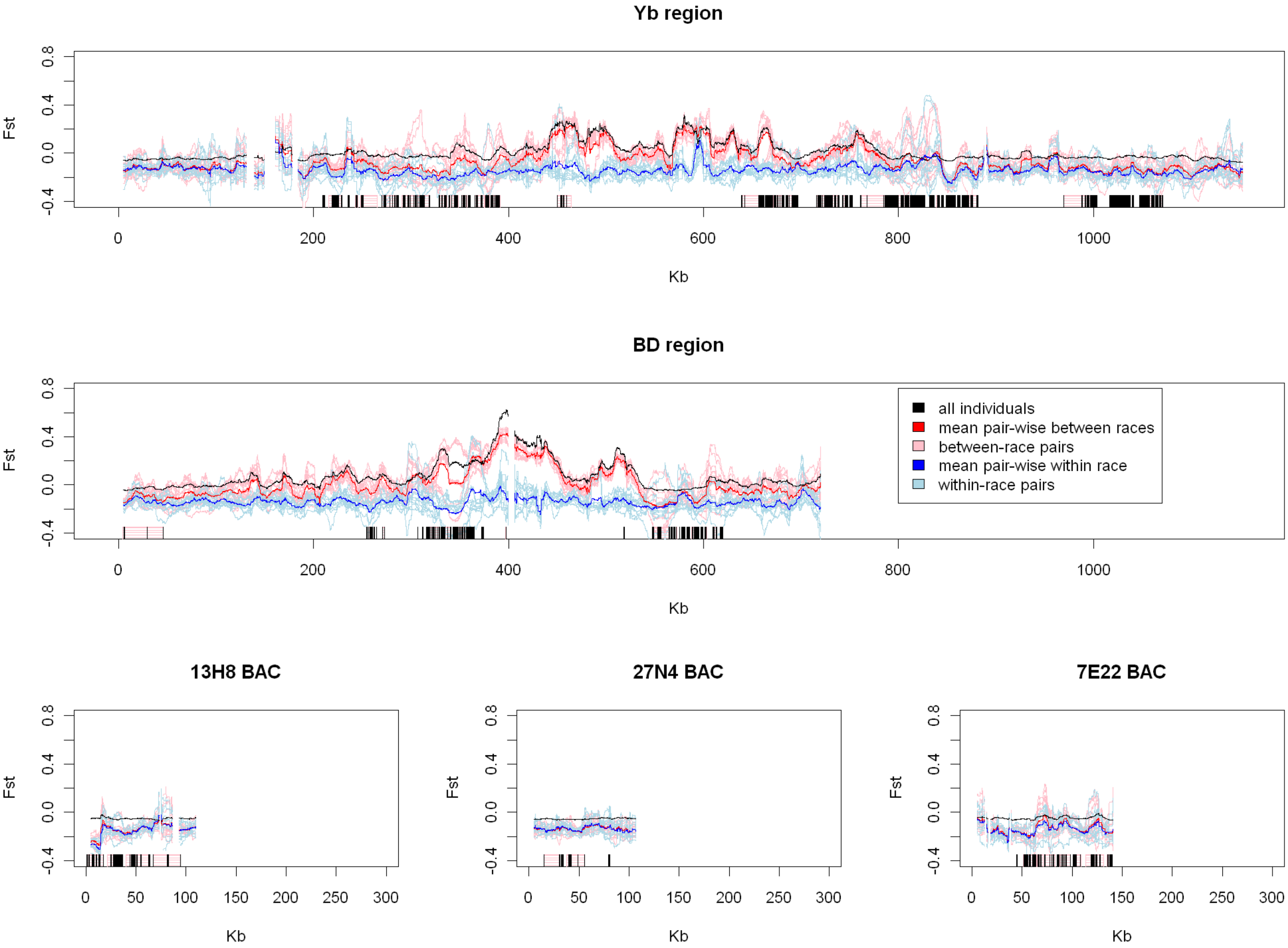


Fig. S1. Pair-wise $F_{ST}$ for all between-race pairs (pink, mean in red) and all within-race pairs (light blue, mean in dark blue) compared to the whole sample estimates (black). Sample size corrections have been applied

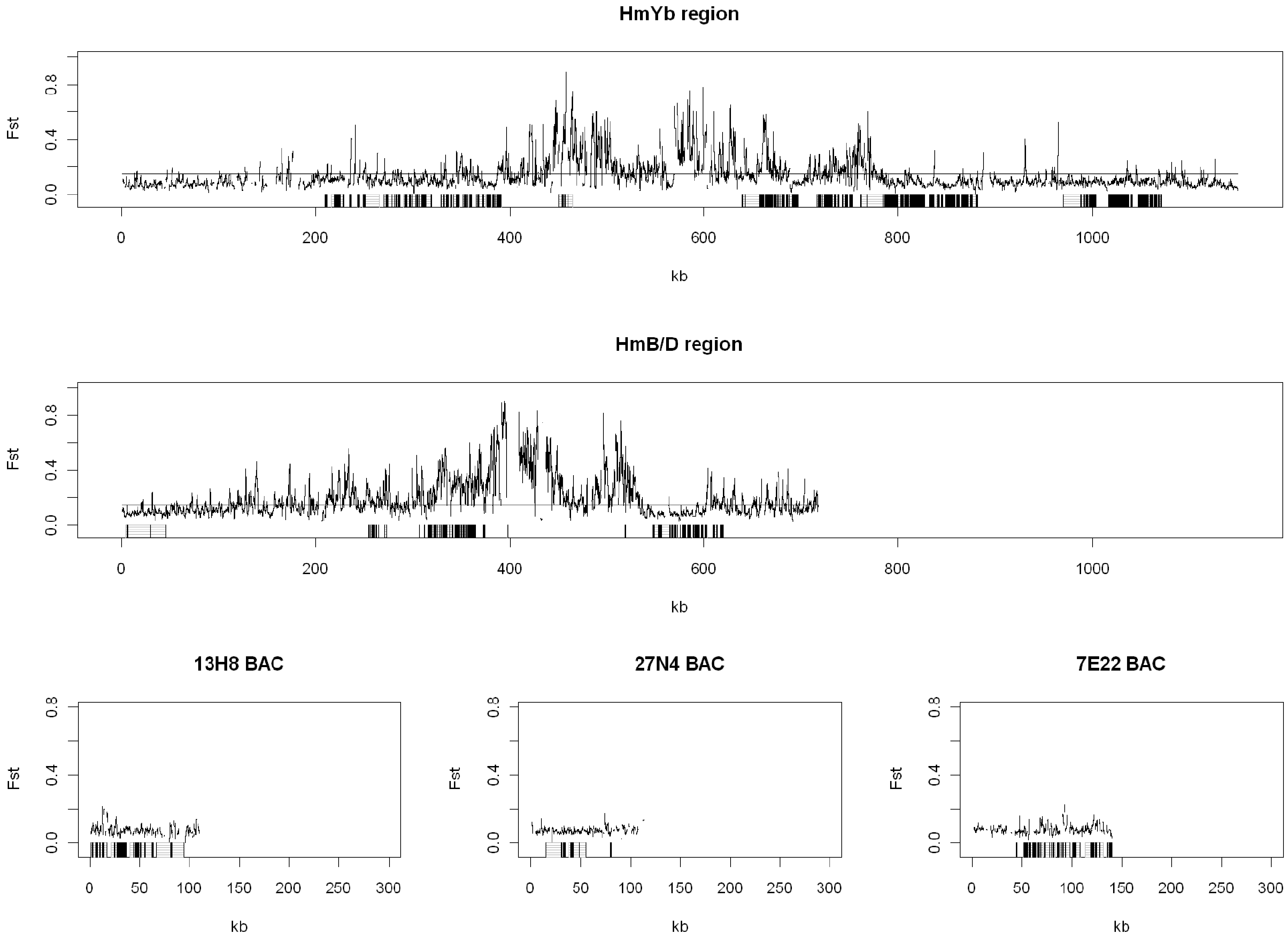


Fig. S2. As figure 2 but with 1 kb windows of $F_{ST}$. Here the threshold in the colour pattern regions is from bootstrap sampling of 100 bps from the unlinked BACs

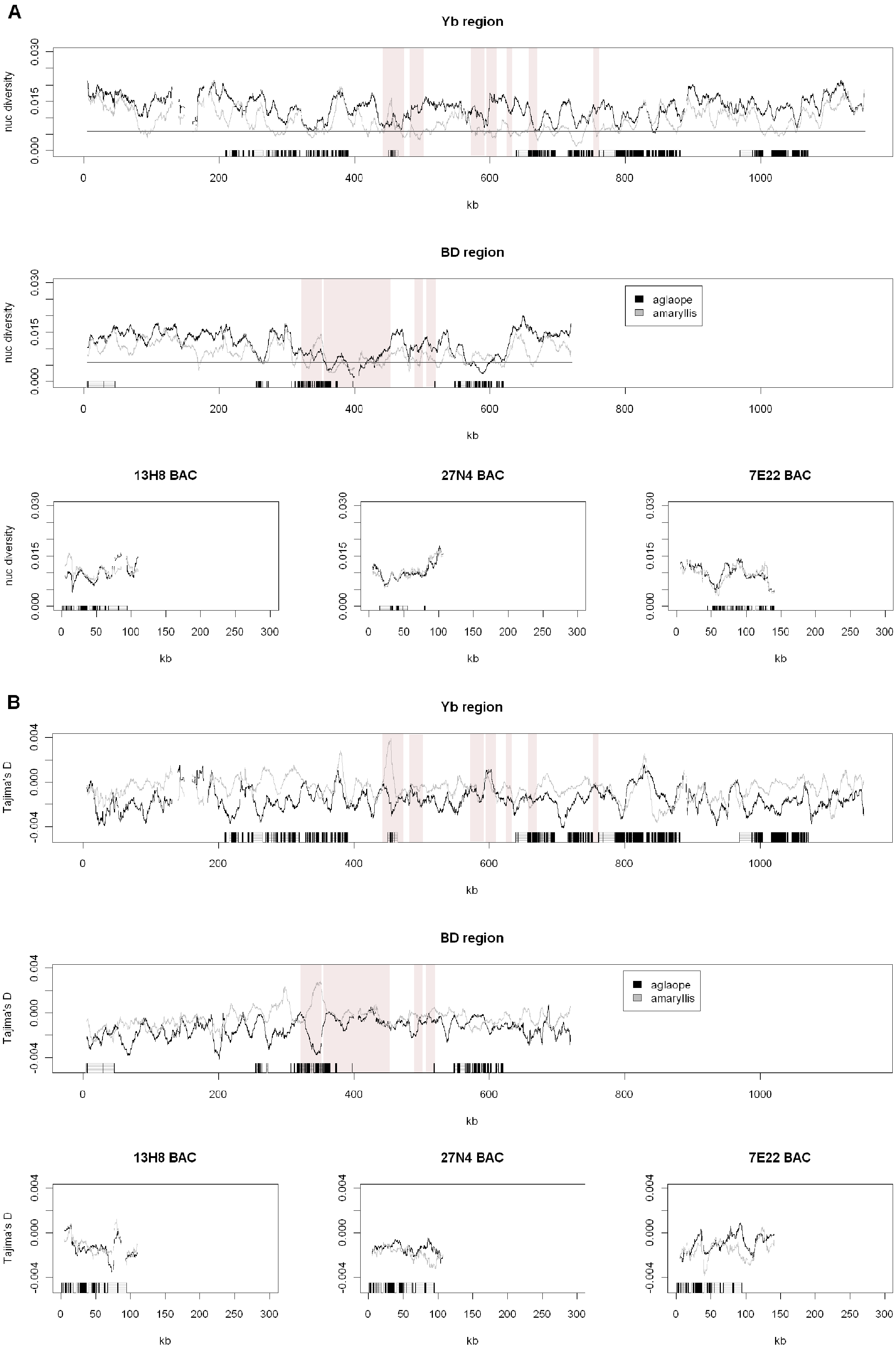


Fig. S3. (a) Nucleotide diversity ($\pi$) and (b) Tajima's *D* across the colour pattern regions and 3 unlinked BACs in *H. melpomene aglaope* (black) and *H. m. amaryllis* (grey). Regions showing peaks of $F_{ST}$ from figure 2 are highlighted in pink.

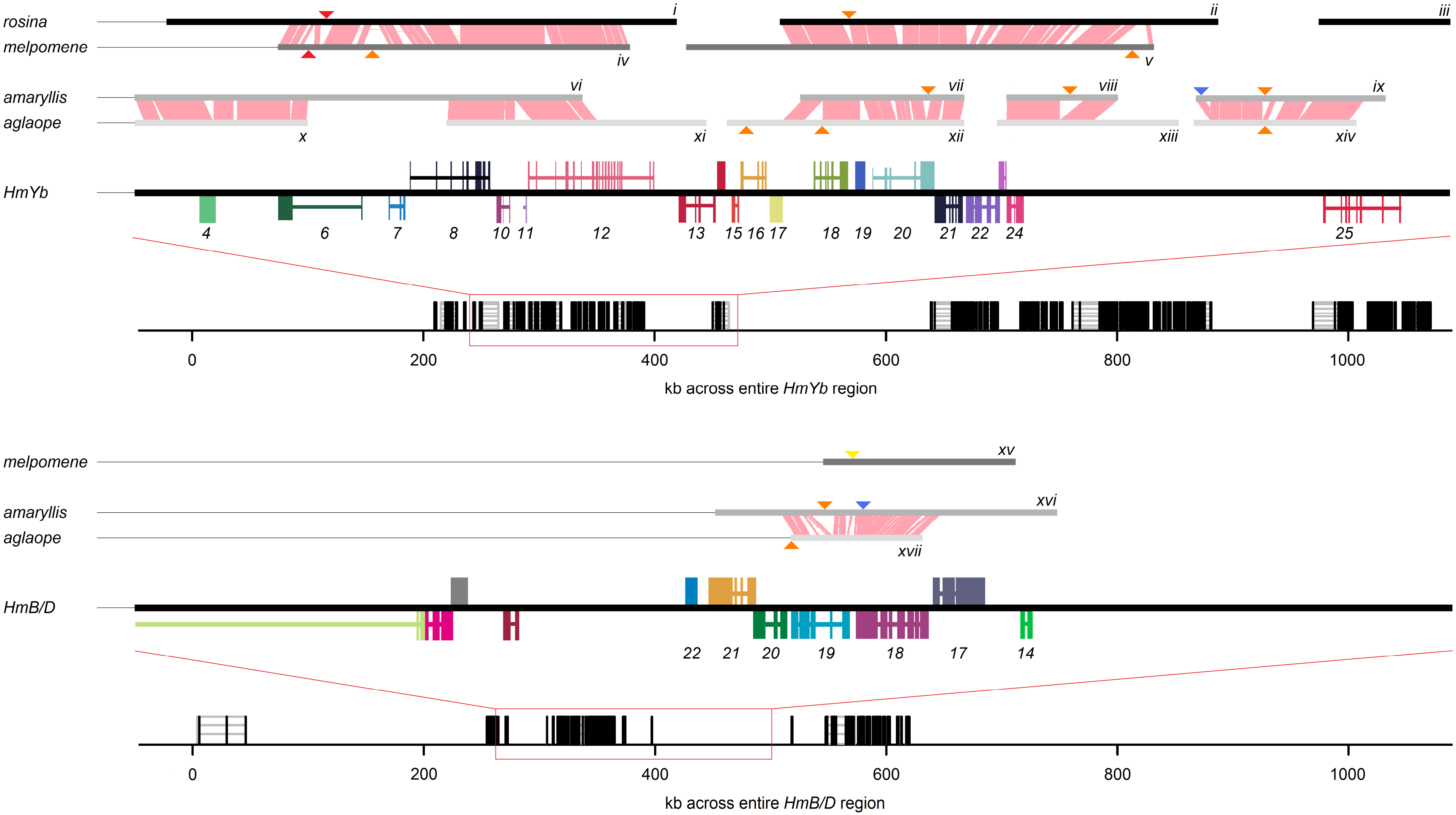


Fig. S4. Sequenced fosmid clone alignments (grey and black bars, alignments in pink). Major differences are indicated with coloured triangles. Gene content and the position relative to the entire *HmYb* and *HmBD* regions are shown below

**Electronic Supplementary Material**

## Supplementary Methods

*Sample preservation and DNA extraction*

For targeted resequencing, adult butterflies were collected, wings removed and bodies preserved in 20% DMSO, 0.2 M EDTA, salt saturated solution. Genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen). For creating the fosmid libraries the four individuals were frozen at -80 $^{o}$C prior to DNA isolation. Tissue was homogenized in DNA isolation buffer (0.1N NaCl, 50mM Tris pH 8.0, 1mM DTT, 10mM EDTA, 0.2% SDS) using a Tissue Lyser (Qiagen) then DNA isolated with three phenol (TE buffered, pH. 8.0) extractions and one chloroform extraction (1). RNA was removed with RNase A (10 mg/ml) and DNA ethanol precipitated and quantified.

*Analysis of capture efficiency and coverage*

For analyses of capture efficiency and coverage we applied a more stringent set of filters. In addition to an upper coverage limit set at twice the median coding sequence coverage for each sample and implemented via Samtools varFilter (2), variant calls with a SNP quality of ≤ 20 or an indel quality of ≤ 50 were also discarded. Summary statistics for the sequencing experiments were obtained using HsMetrics in Picard (version 1.36; http://picard.sourceforge.net). Interval and coverage analyses of the data were performed with the aid of BEDTools (3) and custom scripts in R (version 2.12).

*Fosmid library preparation and sequencing*

Fosmid libraries were prepared using vector pCC1Fos, carrying resistance to chloramphenicol (4). Individual colonies were picked and grown in 384 well plates. Each library contained 206 plates (79,104 clones) that were gridded onto nylon membranes for hybridization.

PCR probes 200-500 bp in length were generated for the genes *HM00004* (*trehalase-1A*), *HM00007* (*B9*)*, HM00010* (*WD40* repeat domain), *HM00013* (*unkempt), HM00017* (*helicase*)*, HM00019* (*BmSuc2 invertase*) *HM00024* (*LRR*) *and HM00025* (*cort*), using *H. melpomene* cDNA as template. These genes have all been annotated in the *HmYb* region, assembled from BAC clones (5). For analysis of part of the *HmB* region, probes were designed for the genes *HM01019* (*Slu7*) *HM01017* (*GPCR*) and *HM01018* (*kinesin-like*). The PCR products were denatured at 95°C (5 min) and labeled with α-32P dCTP using the Prime-a-Gene Labeling System (Promega). Hybridisation of the nylon membranes was performed as outlined in (6). Fosmids were individually subcloned into a sequencing vector (pUC19), sequenced using ABI3730 technology (Applied Biosystems) and traces assembled de novo into contiguous sequences (7).

## Supplementary Results

### *SureSelect targeted resequencing efficiency and coverage of target regions*

In order to evaluate the targeted resequencing experiment, we performed a fine-scale analysis of coverage on the colour pattern regions, summarised in Table S3. Across the 1.4 Mb targeted colour-pattern sequence, the median coverage per sample ranged from 19-62 reads per base. The number of reads that could be mapped back to the reference sequences and the median read depth were both strongly correlated with the amount of sequence data obtained (figure S5). A contribution of repetitive elements within the captured sequence was also evident: localised regions with extremely elevated coverage values were seen. These regions account for the discrepancy between the median and mean coverage values, which was more pronounced in non-coding compared to coding regions. We assessed that regions with ≥200 fold coverage were always due to repeats but to account for differences in the level of coverage between samples we applied a threshold of 2 times the mean coverage of the coding regions. Using this threshold, 10% of targeted sequence was excluded as repetitive. These repeats accounted for over 85% of all read bases within the targeted regions, showing that the repeat library we used for masking prior to bait design was incomplete. Overall, an average of 79% of targeted bases in *H. melpomene* and *H. timareta* and 57% of targeted bases in *H. numata* were retained for downstream analysis. The performance was markedly better in coding regions, where over 97% of the 126 kb of coding bases had adequate coverage in all samples.

To further explore targeted resequencing efficiency, we considered whether characteristics of the baits were related to the depth of coverage we obtained. For all samples, maximal coverage was obtained for baits within the 35-55% GC range (figure S6a). However, this bias is common to all Illumina sequencing, and is attributed to bias in the PCR stages of library preparation, and therefore probably has little to do with the targeting efficiency of the baits (8-10). We also observed an inverse correlation between read depth per bait and the number of single nucleotide mismatches between the bait sequence and the sample (figure S6b). To distinguish between an effect of sequence mismatch on targeted resequencing efficiency versus alignment bias, we applied the strategy of Heap et al. (11) and compared read depth in alignments mapped to the strict reference to read depth in alignments in which mapping bias is reduced by the use of a redundant (consensus) reference sequence into which SNPs identified from the data have been incorporated as ambiguous bases. With the redundant reference, the reduction in read depth with increasing number of SNPs persisted but the impact of sequence mismatch was reduced, particularly as the number of SNPs per bait increased (figure S7). The use of a redundant reference sequence was found to improve the overall coverage, increasing the number of targeted bases with adequate coverage by an average of 2.5% in *H. melpomene* and *H. timareta* samples and 4.2% in *H. numata*.

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

The fosmid sequences were first aligned against our *HmYb* and *HmB/D* BAC reference sequences and then compared against each other (figure S4). In the analysis of the *HmYb* region, *H. m. melpomene* was compared to *H. m. rosina* as these share a hybrid zone in Panama and similarly *H. m. aglaope* was compared to *H. m. amaryllis*. Only *H. m. melpomene*, *H. m. aglaope* and *H. m. amaryllis* were sequenced for the *HmB/D* region, because the Panamanian races share a common red band forewing phenotype.

Within the *HmYb* region, alignment of *H. m. melpomene* and *H. m. rosina* shows sequence shuffling in a 4 kb region upstream of *HM00006* (*trehalase-1A*), within the predicted 5'UTR of the gene. Further upstream in *H. m. melpomene* a retropepsin-like domain and a LTR reverse transcriptase were found, indicative of a transposable element insertion. Similarly, a small reverse transcriptase-like sequence was identified close to the start codon of *HM00019* (*BmSuc2*) of *H. m. rosina*. Indels of approximately 1-2 kb were found in the intronic regions of *HM00008* and *HM00012* of *H. m. rosina* and *H. m. melpomene,* respectively. In this comparison the arrangement differences closest to the regions of high $F_{ST}$ were a series of small indels (20-600 bp in size) in the intergenic region between *HM00024* and *HM00025*.

We found a large transposable element insertion within the second intron of *HM00018 of H. m. aglaope,* which was absent from the *H. m. amaryllis* individual sequenced. There was also evidence of mobile element insertion unique to *H. m. amaryllis*, within the 3'UTR of *HM00020*. Interestingly, we found that the terminal seven exons of gene *HM00025* (*cort*) were in a region of 98% sequence identity between the two species; however, the remaining sequences upstream of these exons, towards the end of the contigs, failed to align, suggesting major sequence differences in the upstream portion of this candidate gene.

Sequencing of the *HmB/D* region was focused around the candidate genes *HM01019 (slu7)*, *HM01018 (kinesin-like)* and *HM01017 (GPCR)*. The most intriguing finding was a duplication event in the sequence of *H. m. amaryllis*: a large part of exon 12 and the whole of the thirteenth (final) exon of the *kinesin-like* gene is repeated approximately 1.2 kb downstream of the complete gene (figure S4). This was unique to *H. m. amaryllis* and not seen in the *H. m. melpomene* or *H. m. aglaope* individuals.

As with the *HmYb* region, we found likely transposable elements in our *HmB/D* fosmid sequences. For example, the *H. m. aglaope* sequence shows evidence of a transposition event between *slu7* and *HM01020*. The two coding regions of these genes are separated by just 830 bp in *H. m. amaryllis*, but the transposition event extends this gap by 3.5 kb, potentially affecting the regulation of either gene. Clearly, further work is needed to determine if any of this structural variation represents fixed differences between races. Although we found no large inversions or structural rearrangements, fosmid sequencing has complemented the targeted resequencing approach, and has identified indels, minor rearrangements and transpositions that differ between the haplotypes studied.

## References

(1) Zraket C, Barth J, Heckel D, Abbott A. 1990 Genetic linkage mapping with restriction fragment length polymorphism in the tobacco budworm, *Heliothis virescens*. In: Molecular Insect Science (ed. Hagedorn HH, Hildebrand JG, Kidwell MG, Law JH). New York: Plenum Press pp. 13-20.

(2) 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.

(3) Quinlan AR, Hall IM. 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841-842.

(4) Quail M, Matthews L, Sims S, Lloyd C, Beasley H, et al. Genomic libraries. I. Construction and screening of fosmid genomic libraries. In: Molecular Methods for Evolutionary Genetics (ed. Orgogozo V, Rockman M). New York: Humana Press; in press.

(5) Ferguson L, Lee SF, Chamberlain N, Nadeau NJ, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, Kee T-J, et al. 2010 Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Molecular Ecology* **19(s1),** 240-254.

(6) Joron M, Papa R, Beltrán M, Chamberlain N, Mavárez J, Baxter S, Ffrench-Constant RH, McMillan WO, Jiggins CD. 2006 A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol.* **4**, e303.

(7) Quail M, Matthews L, Sims S, Lloyd C, Beasley H, Baxter SW. Genomic libraries. II. Subcloning, sequencing, and assembling large-insert genomic DNA clones. In: Molecular Methods for Evolutionary Genetics (ed. Orgogozo, V, and Rockman, M). New York: Humana Press; in press.

(8) Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010 Target-enrichment strategies for next-generation sequencing. *Nat Meth.* **7**, 111-118.

(9) Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. 2008 Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* **5**, 183-188.

(10) Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. 2009 Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**, R32.

(11) Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, Bockett N, et al. 2010 Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet.* **19**, 122-134.

| Clone ID | H.m. linkage group (colour locus) | Length ( bp) | No. of baits | GenBank Accession |
|---|---|---|---|---|
| AEHM-46m10 | LG15 (*HmYb/Sb*) | 112576 | 1176 | CT573313 |
| AEHM-41c10 | LG15 (*HmYb/Sb*) | 118026 | 1273 | CR974474 |
| AEHM-7g12 | LG15 (*HmYb/Sb*) | 118137 | 1253 | CT955980 |
| AEHM-11j7 | LG15 (*HmYb/Sb*) | 78346 | 886 | CU367882 |
| AEHM-29b7 | LG15 (*HmYb/Sb*) | 102012 | 1125 | CU463862 |
| AEHM-21b20 | LG15 (*HmYb/Sb*) | 119880 | 1005 | FP236845 |
| AEHM-22a15 | LG15 (*HmYb/Sb*) | 98783 | 743 | FP245488 |
| AEHM-24o2 | LG15 (*HmYb/Sb*) | 167394 | 1710 | FP102339 |
| AEHM-31b4 | LG15 (*HmYb/Sb*) | 188311 | 1930 | FP102340 |
| AEHM-31j7 | LG15 (*HmYb/Sb*) | 210401 | 2076 | FP102341 |
| AEHM-3o10 | LG15 (*HmYb/Sb*) | 106998 | 1158 | FP236798 |
| AEHM-7g5 | LG15 (*HmYb/Sb*) | 179057 | 2090 | CU462858 |
| AEHM-27i5 | LG18 (*HmB/D*) | 126643 | 1523 | CU467807 |
| AEHM-28l23 | LG18 (*HmB/D*) | 125911 | 1556 | CU467808 |
| AEHM-19l14 | LG18 (*HmB/D*) | 136956 | 1634 | CU672261 |
| AEHM-21p16 | LG18 (*HmB/D*) | 129716 | 1597 | CU681835 |
| AEHM-28f19 | LG18 (*HmB/D*) | 122304 | 1486 | CU672275 |
| AEHM-22C5 | LG18 (*HmB/D*) | 160673 | 1890 | CU462842 |
| AEHM-13H8 | LG20 | 109700 | 1269 | CU525306 |
| AEHM-27N4 | LG1 | 121196 | 1450 | CU468009 |
| AEHM-7E22 | LG13 | 140292 | 1502 | CU856076 |
| | Totals: | 2773312 | 30332 | |

**Table S1**. BAC clones targeted for sequence enrichment

| | HmYb | HmB/D | hm13H8 | hm27N4 | hm7E22 |
|---|---|---|---|---|---|
| $F_{ST}$ (*H. m. aglaope*/*H. m. amaryllis*) | 0.132 (±.002) | 0.184 (±.003) | 0.064 (±.003) | 0.059 (±.002) | 0.065 (±.003) |
| *H. m. aglaope* $\pi$ (%) | 1.27 (±.02) | 1.13 (±.02) | 0.93 (±.06) | 0.91 (±.05) | 0.93 (±.06) |
| *H. m. amaryllis* $\pi$ (%) | 0.87 (±.02) | 0.89 (±.02) | 1.03 (±.07) | 0.98 (±.06) | 0.92 (±.06) |
| *H. m. aglaope* Tajima's *D* | -0.0016 | -0.0014 | -0.0013 | -0.0014 | -0.0010 |
| *H. m. amaryllis* Tajima's *D* | -0.0006 | -0.0005 | -0.0012 | -0.0017 | -0.0016 |

**Table S2.** Population genetics parameters across the colour pattern regions (*HmYb* and *HmB/D*) and across the 3 unlinked BAC sequences (hm13H8, hm27N4,

hm7E22). Values are means for the whole of each of the regions ± 95% confidence intervals calculated from 1000 bootstrap resampling replicates.

| | **Number of sequence reads** | **Mean coverage per base** | | | **Median coverage per base** | | | **% of target with < 10 reads per base** | | | **% of target excluded as repetitive** | | | **% of target with adequate coverage** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **All** | **Coding** | **Non-coding** | **All** | **Coding** | **Non-coding** | **All** | **Coding** | **Non-coding** | **All** | **Coding** | **Non-coding** | **All** | **Coding** | **Non-coding** |
| ***H. m. aglaope*** 09-246 | 11,534,992 | **172.7** | *42.4* | *185.6* | **27** | *42* | *25* | **18.0** | *2.7* | *19.6* | **10.3** | *1.3* | *11.2* | **71.7** | *95.9* | *69.3* |
| ***H. m. aglaope*** 09-267 | 21,893,084 | **351.3** | *85.8* | *377.6* | **56** | *86* | *53* | **10.3** | *0.7* | *11.2* | **10.2** | *0.8* | *11.2* | **79.5** | *98.5* | *77.6* |
| ***H. m. aglaope*** 09-268 | 20,586,148 | **324** | *79* | *348.3* | **53** | *79* | *50* | **9.4** | *0.7* | *10.3* | **10.3** | *0.8* | *11.2* | **80.3** | *98.5* | *78.5* |
| ***H. m. aglaope*** 09-357 | 14,877,258 | **242.8** | *61.2* | *260.8* | **37** | *60* | *34* | **13.7** | *1.0* | *15.0* | **10.2** | *0.9* | *11.1* | **76.1** | *98.1* | *73.9* |
| ***H. m. amaryllis s***09-332 | 21,327,952 | **343.2** | *78.6* | *369.5* | **51** | *79* | *48* | **11.1** | *0.9* | *12.1* | **10.2** | *0.7* | *11.2* | **78.7** | *98.4* | *76.8* |
| ***H. m. amaryllis*** 09-333 | 22,507,694 | **344.7** | *79.4* | *371* | **51** | *79* | *48* | **10.7** | *0.8* | *11.7* | **10.4** | *0.8* | *11.3* | **78.9** | *98.4* | *77.0* |
| ***H. m. amaryllis*** 09- 75 | 12,930,662 | **201.6** | *52.2* | *216.4* | **34** | *51* | *32* | **14.0** | *1.7* | *15.3* | **10.2** | *1.1* | *11.1* | **75.8** | *97.1* | *73.6* |
| ***H. m. amaryllis*** 09- 79 | 24,836,302 | **399.4** | *96.9* | *429.3* | **62** | *97* | *59* | **8.8** | *0.5* | *9.7* | **10.4** | *0.7* | *11.3* | **80.8** | *98.8* | *79.0* |
| ***H. timareta ssp. Nov.*** | 15,135,324 | **245.4** | *61.2* | *263.7* | **39** | *62* | *37* | **11.7** | *1.3* | *12.8* | **10.3** | *0.9* | *11.2* | **78.0** | *97.8* | *76.0* |
| ***H. numata*** | 9,697,212 | **113.4** | *45.1* | *120.6* | **19** | *44* | *16* | **34.2** | *3.7* | *37.2* | **9.1** | *1.1* | *9.9* | **56.7** | *95.2* | *52.9* |

**Table S3.** Coverage of the colour pattern regions across individual samples. Repeats are defined here as regions with > 2 times the median coverage of the coding regions. "% target with adequate coverage" are those regions that fall below this maximum threshold and above the minimum threshold.

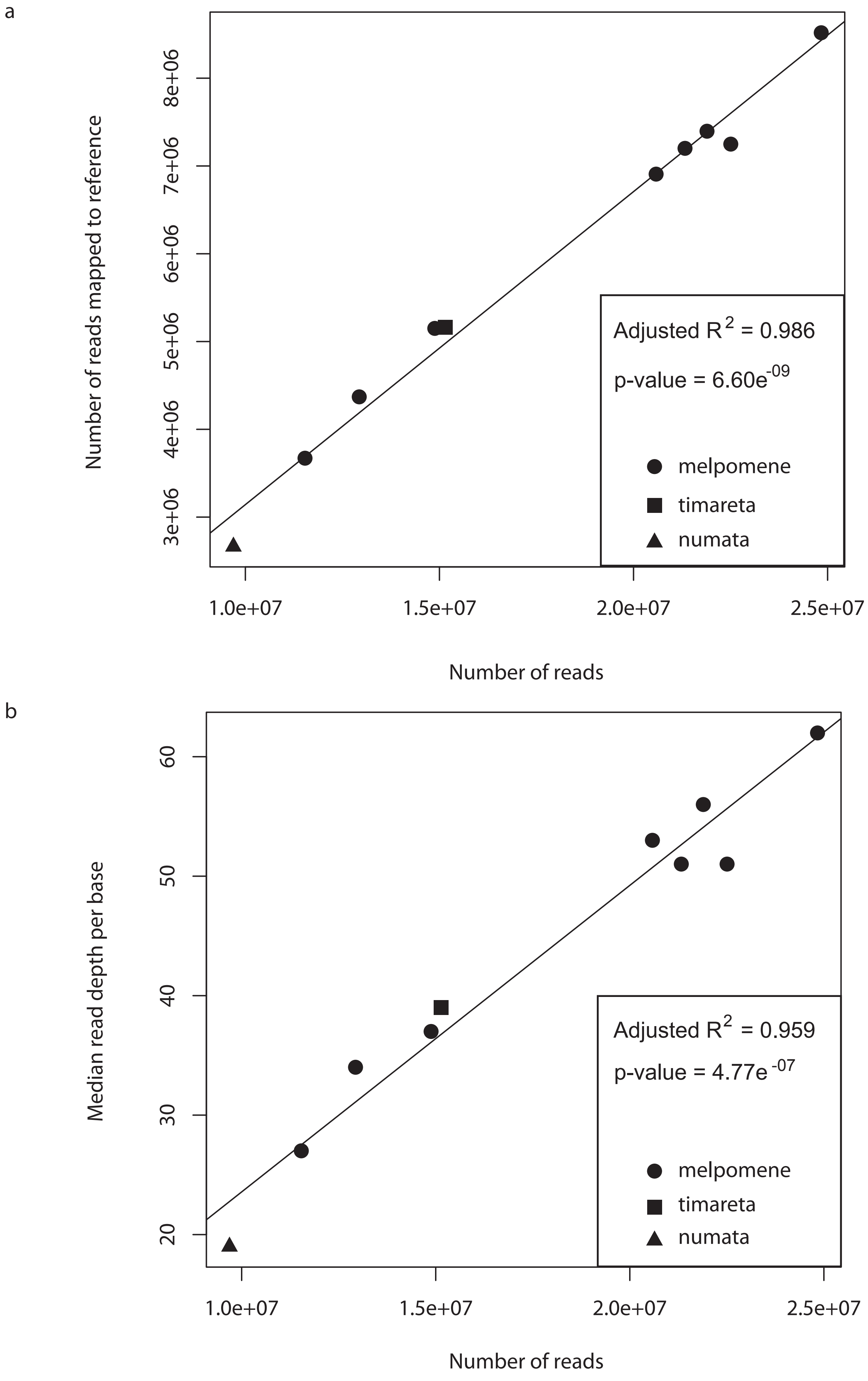

a
Number of reads mapped to reference
3e+06
4e+06
5e+06
6e+06
7e+06
8e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
Number of reads
Adjusted $R^2$ = 0.986
p-value = $6.60e^{-09}$
● melpomene
■ timareta
▲ numata
b
Median read depth per base
20
30
40
50
60
1.0e+07
1.5e+07
2.0e+07
2.5e+07
Number of reads
Adjusted $R^2$ = 0.959
p-value = $4.77e^{-07}$
● melpomene
■ timareta
▲ numata

**Figure S5.** Relationships between total number of reads per sample and a) the number of reads mapping back to the reference sequence and b) the median read depth within the colour pattern regions.

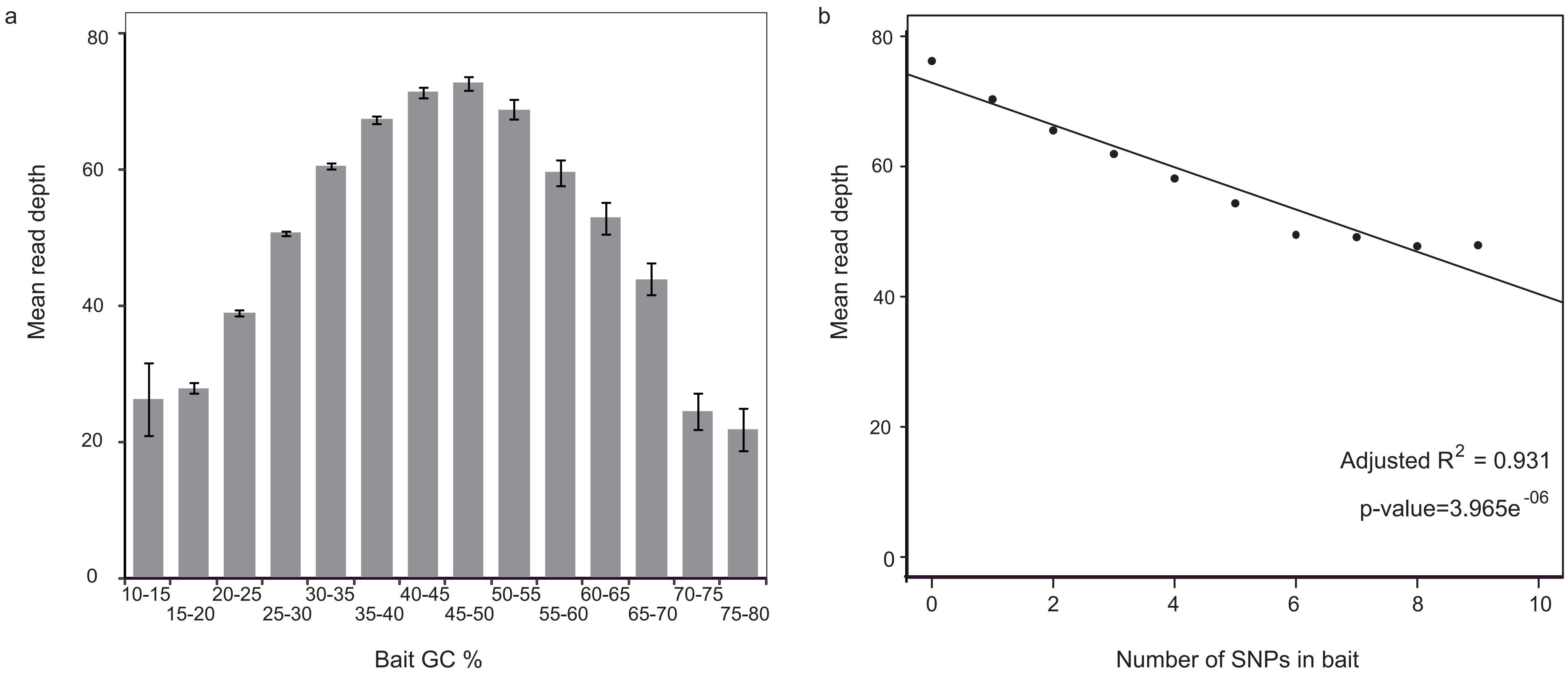
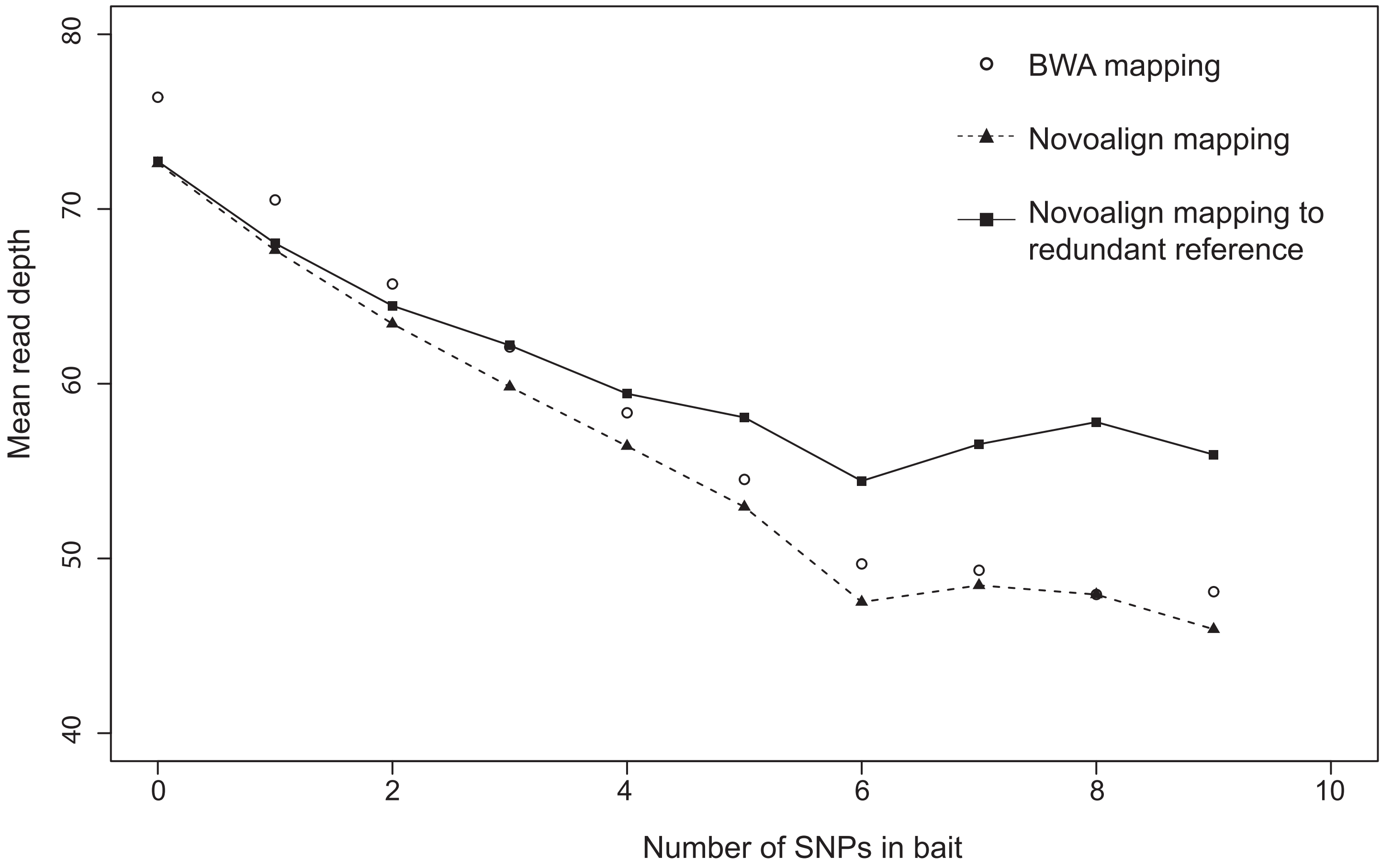

a
80
60
40
20
0
Mean read depth
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
Bait GC %
b
80
60
40
20
0
Mean read depth
Adjusted R$^2$ = 0.931
p-value=3.965e$^{-06}$
0
2
4
6
8
10
Number of SNPs in bait

**Figure S6:** The effect of GC content and number of sequence variants on the number of reads mapping. The number of SNPs (single nucleotide polymorphisms) refers to the number of nucleotides that differ from the reference. Data are from one individual (*H. m. aglaope* 09-267) but results were very similar for all individuals.



80
70
60
50
40
Mean read depth
0
2
4
6
8
10
Number of SNPs in bait
BWA mapping
Novoalign mapping
Novoalign mapping to redundant reference

**Figure S7**. Comparison of mapping strategies. Alignment to a redundant reference which includes identified SNP variation improves mapping performance, particularly of divergent reads. Data are from one individual (*H. m. aglaope* 09-267) but results for all individuals were very similar.