Published in "Nature: 10.1038/nature10341, 2011"
which should be cited to refer to this work.



# Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry

Mathieu Joron[1,2,3], Lise Frezal[1]*, Robert T. Jones[4]*, Nicola L. Chamberlain[4], Siu F. Lee[5], Christoph R. Haag[6], Annabel Whibley[1], Michel Becuwe[2], Simon W. Baxter[7], Laura Ferguson[7], Paul A. Wilkinson[4], Camilo Salazar[8], Claire Davidson[9], Richard Clark[9], Michael A. Quail[9], Helen Beasley[9], Rebecca Glithero[9], Christine Lloyd[9], Sarah Sims[9], Matthew C. Jones[9], Jane Rogers[9], Chris D. Jiggins[7] & Richard H. ffrench-Constant[4]

**Supergenes are tight clusters of loci that facilitate the co-segregation of adaptive variation, providing integrated control of complex adaptive phenotypes[1]. Polymorphic supergenes, in which specific combinations of traits are maintained within a single population, were first described for 'pin' and 'thrum' floral types in *Primula*[1] and *Fagopyrum*[2], but classic examples are also found in insect mimicry[3–5] and snail morphology[6]. Understanding the evolutionary mechanisms that generate these co-adapted gene sets, as well as the mode of limiting the production of unfit recombinant forms, remains a substantial challenge[7–10]. Here we show that individual wing-pattern morphs in the polymorphic mimetic butterfly *Heliconius numata* are associated with different genomic rearrangements at the supergene locus *P*. These rearrangements tighten the genetic linkage between at least two colour-pattern loci that are known to recombine in closely related species[9–11], with complete suppression of recombination being observed in experimental crosses across a 400-kilobase interval containing at least 18 genes. In natural populations, notable patterns of linkage disequilibrium (LD) are observed across the entire *P* region. The resulting divergent haplotype clades and inversion breakpoints are found in complete association with wing-pattern morphs. Our results indicate that allelic combinations at known wing-patterning loci have become locked together in a polymorphic rearrangement at the *P* locus, forming a supergene that acts as a simple switch between complex adaptive phenotypes found in sympatry. These findings highlight how genomic rearrangements can have a central role in the coexistence of adaptive phenotypes involving several genes acting in concert, by locally limiting recombination and gene flow.**

The origin and maintenance of adaptive multi-locus polymorphism in the face of recombination is a long-standing puzzle in evolutionary biology[7,12,13]. In some cases, supergene architecture has evolved with tight linkage that maintains specific combinations of alleles at neighbouring genes[1–6]. A notable illustration is provided by polymorphic mimetic butterflies, in which several discrete forms, each resembling a different model, are maintained in sympatry. Examples include Batesian polymorphism in *Papilio dardanus*[5,14] and *Papilio memnon*[3] and Müllerian polymorphism in the neotropical species *Heliconius numata*[4,15]. In each case, a single supergene locus controls coordinated differences in a complex phenotype which can involve modifications of wing pattern and shape, body colour and perhaps behaviour[3,4]. Mimetic patterns represent sharp fitness peaks corresponding to locally abundant wing patterns, separated by adaptive valleys in which selection acts against recombinant individuals with intermediate, non-mimetic phenotypes[16].

Theoretical debate has centred on the constraints imposed on supergene evolution by genomic organization, specifically whether loci must be tightly linked from the outset or whether the association between elements can be acquired, either gradually or in a single mutational step[7–10,16–19]. Chromosomal rearrangements, which can bring genes into closer physical association and influence local recombination, offer one route through which supergenes may be assembled from more loosely linked components[7,8,17–19]. Although there are many cases of polymorphic inversions associated with adaptive variation[12,20], variation is in most cases geographical, rather than being maintained within populations, and effects on local adaptation are cumulative. In contrast, supergenes are characterized by a Mendelian switch between clearly defined combinations of traits in populations. Here we investigate the genomic organization and population genetics of the *P* supergene in *H. numata*[4,9] and identify a key role for structural variation in strengthening and maintaining allelic associations within the supergene.

In *H. numata*, up to seven sympatric morphs coexist in local populations and each is an accurate mimic of one of several available model species in another butterfly family (Danainae: *Melinaea*)[4,9]. Each morph is controlled by a specific allele at *P* with precise allelic dominance[4,15]

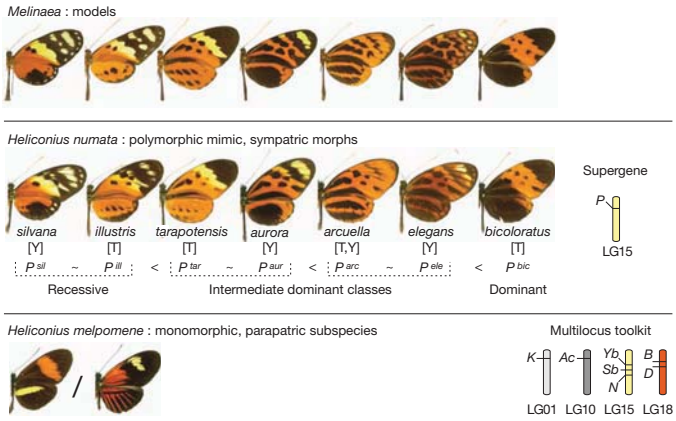


**Figure 1 | Supergene alleles and mimicry polymorphism in *H. numata*.** Polymorphic forms of *H. numata* each mimic different models in the distantly related genus *Melinaea* (Nymphalidae: Danainae). Each form is controlled by a specific allele of the supergene *P*, with increasing dominance shown from left to right[4,9]. Two parapatric regions of northeastern Peru (T, Tarapoto and Andean valleys; Y, Yurimaguas and Amazon lowlands) harbour different mimicry assemblages[15]; dominance (<) is nearly complete between forms within each region, but is incomplete (~) between certain pairs of alleles from parapatric regions. In all other species studied in the genus *Heliconius*, wing pattern is controlled by several large-effect loci on different chromosomes. In *H. melpomene* the *HmYb–HmSb–HmN* complex is situated in the orthologous position to the *H. numata P* supergene[9]. LG, linkage group.

[1]CNRS UMR 7205, Muséum National d'Histoire Naturelle, CP50, 45 Rue Buffon, 75005 Paris, France. [2]Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK. [3]Institute of Biology, Leiden University, Postbus 9505, 2300 RA Leiden, The Netherlands. [4]Centre for Ecology and Conservation, School of Biosciences, University of Exeter, Cornwall Campus, Penryn, Cornwall TR10 9EZ, UK. [5]Department of Genetics, Bio21 Institute, University of Melbourne, 30 Flemington Road, Parkville, 3010 Victoria, Australia. [6]Department of Biology, Ecology and Evolution, University of Fribourg, Chemin du Musée 10, CH-1700 Fribourg, Switzerland. [7]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. [8]Smithsonian Tropical Research Institute, NAOS island, Causeway Amador, Panamá, República de Panamá. [9]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK.
*These authors contributed equally to this work.

(Fig. 1). Rare non-mimetic individuals that combine pattern elements from different morphs are observed at a frequency of <0.7% in natural polymorphic populations[4,15]. These individuals, presumed to be recombinants, confirm the existence of several functional elements in the locus. The genomic position of *P* was previously shown to correspond to a cluster of three loci, *HmN*, *HmYb* and *HmSb*, on linkage group 15 in the closely related species *Heliconius melpomene*[9], in which they control distinct wing-pattern elements in geographical races. Other unlinked wing-pattern loci in related species do not have large effects on *H. numata* mimicry[9,21].

By fine-scale linkage mapping and positional cloning, we identified a chromosomal interval of about 400 kilobases (kb) containing the *H. numata P* supergene, defined by the absence of crossing over in 366 progeny from six broods (Fig. 2a). The orthologous region in *H. melpomene* shows notable rates of recombination and contains two distinct colour-pattern loci, *HmYb* and *HmSb*, ~0.9 centimorgans (cM) apart[11]. Although genome-wide estimates of recombination rate in *Heliconius* are not available, markers adjacent to *P*, markers elsewhere on linkage group 15, and markers in other linkage groups (Supplementary Table 1)[9] all show significantly higher rates of crossing over (chi-squared test of independence, $P = 3.8 \times 10^{-6}$), consistent with severe suppression of recombination at *P* in *H. numata*.

To examine the gene content and genomic organization of the supergene, we sequenced an approximately one-megabase (Mb) region centred on *P* by screening a bacterial artificial chromosome (BAC) library prepared from a mixture of individuals from a single polymorphic population of *H. numata* (Supplementary Table 2). Colinearity and gene content was generally conserved with *H. melpomene*[11], but two BACs overlapping the *P* interval showed distinct gene orders, both different from the order seen in *H. melpomene* (Figs 2a, 3 and Supplementary Fig. 1). A third gene order, corresponding to the *H. melpomene* reference, was detected by PCR (see below). Gene orientation near the breakpoints on the *H. numata* clones indicated a minimum of two rearrangement events compared to the reference, putatively involving 31 genes: a 400-kilobase (kb) segment containing 18 genes from *HN00023* to *HN00040* (breakpoint BP1), and a 180-kb segment containing 13 genes from *HN00041* to *HN00053* (breakpoint BP2) (Fig. 3a and Supplementary Fig. 1). The altered gene orders at *P* contrast with the colinearity of flanking sequences between *H. numata* and *H. melpomene* (Supplementary Fig. 2 and Supplementary Tables 3 and 4), as well as between more distantly related *Heliconius* species[22]. Furthermore, *H. melpomene* and the silk moth *Bombyx mori*, separated by about 100 million years of evolution, share a generally conserved gene order across this region[11,23]. This supports the hypothesis that the

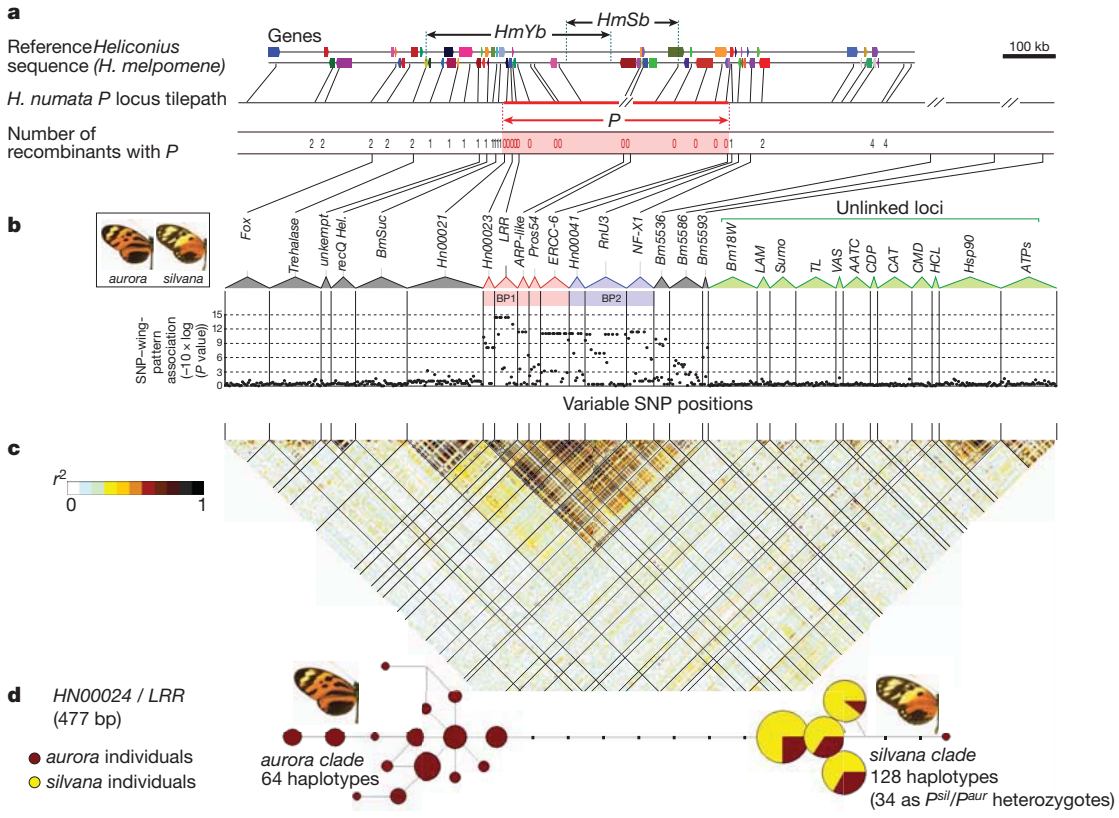


**Figure 2 | Fine-scale mapping and nucleotide variation at the *P* supergene in *H. numata*.** **a**, Fine-scale linkage mapping of the supergene *P* (indicated by red arrows) to the interval bounded by genes *HN00020* and *HN00041*. Recombinants were observed in crosses totalling 366 individuals. Blue arrows indicate the position of the recombining loci *HmYb* and *HmSb* in *H. melpomene*[11]. *HmN*, which is known to be part of this cluster of loci, is not fine-mapped in *H. melpomene*[9,11]. Coloured blocks represent annotated gene regions on forward and reverse strands (see Supplementary Fig. 1 and Supplementary Table 3 for details). **b**, Association of SNP variation with mimicry polymorphism in a sample of 25 *silvana* and 34 *aurora*/*arcuella* individuals from a single population. Markers genotyped across rearrangements BP1 (pink) and BP2 (blue) show perfect association of SNP variation with wing pattern. No association was found in the flanking region from markers *Fox* (*Hn00106*) to *BmSuc* (*Hn00019*), or at 12 unlinked loci (green). The association decays more slowly in the direction of loci *Bm5536* (*GCP*), *Bm5586* (*NudC*) and *Bm5593* (*Srp68*). **c**, LD heat map. Perfect LD (genotypic correlation coefficient $r^2 = 1$) is found across 580 kb spanning the BP1 and BP2 rearrangements ($n = 59$). LD decays rapidly outside this interval, although strong within-marker LD remains at *HN00021*. Markers that are unlinked to *P* show little LD with each other or with *P*. **d**, Haplotype network for marker *LRR* (*Hn00024*) in the Yurimaguas population, coloured according to wing-pattern phenotype. Haplotype clades separated by seven fixed differences are in complete association with wing pattern, taking into account dominance relationships. Similar haplotype clades were found for all loci genotyped within *P*, and across the Amazon basin, but not for genes flanking *P* or in unlinked regions (Supplementary Figs 4 and 6).

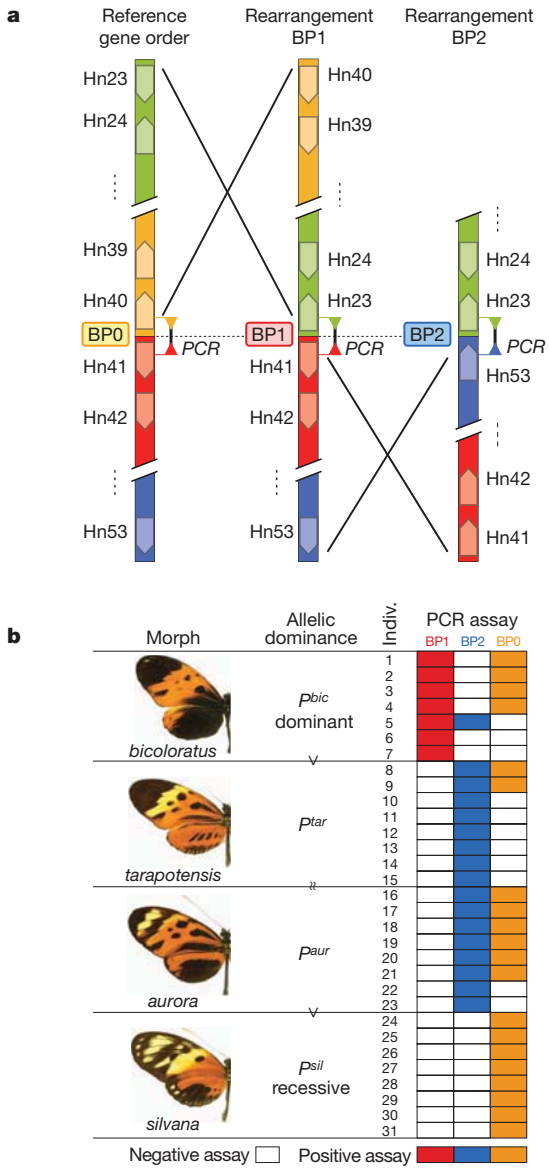


**Figure 3 | Chromosomal rearrangements associated with the supergene in natural populations.** **a**, Comparison of the gene orders found in the *H. numata* BAC library and wild populations. The rearrangements involve the 400-kb segment from genes *HM00023* to *HM00040* (*ERCC6*) (BP1, clones 24I10 and 45B17), and the adjacent 180-kb segment from gene *HN00041* (*penguin*) and *HN00053* (*lethal (2) giant larvae homologue*) (BP2, clone 38G4). Genes closest to the breakpoints are shown. **b**, Long-range PCR assays across alternative breakpoints (BP0, BP1 and BP2) in wild populations show a perfect association of the polymorphic gene orders with mimicry variation in four morphs from natural populations of Eastern Peru[15], following the dominance relationships (Fig. 1). The Yurimaguas population segregates primarily for *silvana* and *aurora*/*arcuella* forms, associated with BP0 and BP2, respectively. The Tarapoto population segregates mainly for *tarapotensis* and *bicoloratus* forms, associated with BP2 and BP1, respectively. This population also harbours recessive *illustris* alleles associated with BP0 (Supplementary Table 6).

gene rearrangements are evolutionarily derived and associated with the evolution of this locus in the *H. numata* lineage.

The results from BAC-clone sequencing were extended to natural populations by performing breakpoint-specific PCR on butterflies collected in the field (Fig. 3a and Supplementary Table 5). Long-range PCR analyses of 31 individuals of four morphs revealed a complete association between alternative rearrangements and specific wing-pattern phenotypes (Fig. 3b). Rearrangement BP1 was found in every *H. numata bicoloratus* and in no other morph, as expected for the dominant *P* allele, whereas BP2 was found in all intermediate dominant forms (*aurora* and *tarapotensis*). All recessive *silvana* individuals were BP0 homozygotes, confirming that the reference gene order segregates in *H. numata* populations. Some individuals with intermediate and dominant alleles had breakpoints that were lower in the dominance series, reflecting their heterozygosity. To confirm the association on a larger sample, short-range PCR assays primed closer to the breakpoints were performed on 156 individuals of six major morphs (Fig. 1 and Supplementary Tables 5–7). Short-range assays were highly consistent with long-range assays (Supplementary Table 6). Natural populations thus harbour at least three chromosomal arrangements in tight association with wing-pattern phenotypes. A short-range PCR product from BP1 was not amplified in two of the 32 *bicoloratus* individuals tested. This could reflect PCR failure due to sequence variation at BP1, or it may indicate that the BP1 breakpoint is not directly causative of the *bicoloratus* phenotype, despite being strongly associated with the causative variant(s).

To assess whether the recombination suppression that is observed in crosses operates at the population level, we estimated LD between nucleotide positions in and around *P*. A notable pattern of long-range LD was seen in tight association with the position of *P*, mapped from a survey of 17 markers sequenced in 59 individuals from the polymorphic population near Yurimaguas (Figs 1 and 2c). Complete LD was found between high-frequency single nucleotide polymorphisms (SNPs) across the 400-kb *P* interval, showing that long-range haplotypes are maintained across the supergene. This haplotype structure decays rapidly outside *P*: flanking markers lack the haplotype segregation found within *P* and show comparable levels of LD to the levels seen in 12 unlinked markers (Fig. 2c and Supplementary Fig. 3). Furthermore, within *P*, complete association was found between the divergent haplotypes and wing-pattern alleles. All recessive *silvana* individuals were homozygous for 39 SNPs in six genes across the *P* locus, whereas *aurora* individuals were heterozygous or homozygous for the alternative nucleotides (Fig. 2b, Supplementary Fig. 4 and Supplementary Table 5). Again, this complete association breaks down immediately outside *P* (Fig. 2b and Supplementary Figs 4 and 5). Therefore, the divergent haplotypes associated with wing-pattern alleles seem to be confined within the boundaries of the polymorphic rearrangements and the position of the supergene.

Finally, to confirm that the haplogroups are associated with phenotype and are not due to local population processes, we screened a population of *H. numata* from French Guiana, situated 2,900 km from Peru but harbouring phenotypically similar morphs (*silvana* and *numata*; Supplementary Table 5)[4]. The same two haplotype groups were found, and breakpoint PCR and haplotype clades were perfectly associated with corresponding phenotypes in both locations (Supplementary Fig. 6 and Supplementary Table 6). The association between mimicry polymorphism, local rearrangements and divergent haplotype blocks therefore seems to be conserved across the Amazonian range.

Our results show that the supergene is characterized by a 400-kb chromosomal block that is sharply structured into distinct haplotype clades separated by 1–4% divergence. These clades correspond to different wing-pattern alleles, segregate consistently with the allelic dominance and are associated with the chromosomal rearrangements (Fig. 2d and Supplementary Figs 3 and 4). This situation stands in stark contrast to the related species *H. melpomene* and *H. erato*, in which several independent wing-patterning loci are under directional selection, and in which markers surveyed in the orthologous region showed no fixed nucleotide differences between colour-pattern races, and no long-range LD[10,24]. The *P* supergene architecture, associated with mimicry polymorphism under balancing selection in *H. numata*[4,15], is thus evolutionarily characterized by non-recombining co-adapted blocks that capture distinct wing-pattern genes that are known to recombine in other species of the clade[10,24].

In summary, the four strands of evidence, namely fine-scale comparative mapping, association of three chromosomal arrangements with morphs, SNP–phenotype association at *P*, and long-range LD

and divergent haplotype clades, all indicate that the rearrangements lock together distinct elements involved in wing-pattern evolution, providing long-awaited evidence for a situation that may apply to the evolution of supergenes in other systems[7,16–19]. Together, the data begin to explain how distinct loci that control geographical variation in some species[10,11] can become locked together in others. Our results highlight the role that structural variation can have in generating co-adapted gene complexes involved in adaptation and speciation[8,12,19,20,25], and open the way to the study of their functional integration.

The next challenge will be to identify the sites that are causally involved in the elements of pattern variation in *H. numata*, a goal hindered by reduced recombination within *P*, which limits the power of fine mapping to dissect the locus. Instead, functional studies of genes in the interval and analysis of rare recombinants will be important. Notably, *P* seems to be a hotspot of adaptation in several other species[26], including the melanic peppered moth[27], and, along with other colour-pattern regions, is linked to assortative mating and speciation in other *Heliconius* species[28–30]. Unravelling the genetic nature of the elements that contribute to the supergene will therefore be key to understanding the mechanisms and sequence of events underlying the clustering of adaptive traits in rearrangements that are associated with ecological divergence.

## METHODS SUMMARY

A BAC library was screened with radiolabelled probes derived from BAC end-sequences during chromosome walking. Clone sequences, obtained by capillary sequencing, were annotated with the aid of a Roche 454-sequenced cDNA library from the developing wing[11]. Genomic sequences were aligned to a *H. melpomene* reference[11] to detect chromosomal rearrangements. Crossover events were fine-mapped using broods described elsewhere[9]. Recombination estimates from other genomic regions were derived by scoring crossing over along two unlinked clones, using markers separated by known physical distance.

PCR-based diagnostics for the rearrangement breakpoints BP1 and BP2 were designed from *H. numata* BAC sequences. BP0 was designed from the reference *H. melpomene* sequence. Long-range PCRs were primed in exons flanking each breakpoint, and verified by sequencing. Short-range PCRs were primed closer to the breakpoints in unique intergenic regions (Supplementary Table 7).

Genetic markers were sequenced by capillary sequencing. LD was estimated from unphased sequences using genotypic correlation tests between all pairs of variable sites. The association between wing-pattern and SNPs was tested using Fisher's exact tests, taking into account the dominance hierarchy of wing-pattern alleles. Haplotype networks were constructed by parsimony using Network.


1. Mather, K. The genetical architecture of heterostyly in *Primula sinensis*. *Evolution* **4,** 340–352 (1950).
2. Garber, R. J. & Quisenberry, K. S. The inheritance of length of style in buckwheat. *J. Agric. Res.* **34,** 181–183 (1927).
3. Clarke, C. A., Sheppard, P. M. & Thornton, I. W. B. The genetics of the mimetic butterfly *Papilio memnon*. *Philos. Trans. R. Soc. Lond. B* **254,** 37–89 (1968).
4. Brown, K. S. & Benson, W. W. Adaptive polymorphism associated with multiple Müllerian mimicry in *Heliconius numata*. *Biotropica* **6,** 205–228 (1974).
5. Nijhout, H. F. Polymorphic mimicry in *Papilio dardanus*: mosaic dominance, big effects, and origins. *Evol. Dev.* **5,** 579–592 (2003).
6. Murray, J. & Clarke, B. Supergenes in polymorphic land snails—examples from genus *Partula*. *Genetics* **74,** S188–S189 (1973).
7. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173,** 419–434 (2006).
8. Manfield, I. W. *et al.* Molecular characterization of DNA sequences from the *Primula vulgaris S*-locus. *J. Exp. Bot.* **56,** 1177–1188 (2005).
9. Joron, M. *et al.* A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biol.* **4,** e303 (2006).
10. Baxter, S. W. *et al.* Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet.* **6,** e1000794 (2010).
11. Ferguson, L. *et al.* 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 (2010).
12. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39,** 21–42 (2008).
13. Pinho, C. & Hey, J. Divergence with gene flow: Models and data. *Annu. Rev. Ecol. Evol. Syst.* **41,** 215–230 (2010).
14. Clark, R. *et al.* Colour pattern specification in the Mocker swallowtail *Papilio dardanus*: the transcription factor invected is a candidate for the mimicry locus H. *Proc. R. Soc. B* **275,** 1181–1188 (2008).
15. Joron, M., Wynne, I. R., Lamas, G. & Mallet, J. Variable selection and the coexistence of multiple mimetic forms of the butterfly *Heliconius numata*. *Evol. Ecol.* **13,** 721–754 (1999).
16. Turner, J. R. G. in *The biology of butterflies* Vol. 11 (eds Vane-Wright, R. I. & Ackery, P. R.) 141–161 (Academic, 1984).
17. Charlesworth, D. & Charlesworth, B. Theoretical genetics of Batesian mimicry. II. Evolution of supergenes. *J. Theor. Biol.* **55,** 305–324 (1975).
18. Alvarez, G. & Zapata, C. Conditions for protected inversion polymorphism under supergene selection. *Genetics* **146,** 717–722 (1997).
19. Hatadani, L. M., Baptista, J. C. R., Souza, W. N. & Klaczko, L. B. Colour polymorphism in *Drosophila mediopunctata*: genetic (chromosomal) analysis and nonrandom association with chromosome inversions. *Heredity* **93,** 525–534 (2004).
20. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8,** e1000500 (2010).
21. Jones, R. T., Salazar, P., ffrench-Constant, R. H., Jiggins, C. D. & Joron, M. Evolution of a mimicry supergene from a multilocus architecture. *Proc. R. Soc. B*. doi:10.1098/rspb.2011.0882 (2011).
22. Papa, R. *et al.* Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in *Heliconius* butterflies. *BMC Genomics* **9,** 345 (2008).
23. Pringle, E. G. *et al.* Synteny and chromosome evolution in the lepidoptera: Evidence from mapping in *Heliconius melpomene*. *Genetics* **177,** 417–426 (2007).
24. Counterman, B. A. *et al.* Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genet.* **6,** e1000796 (2010).
25. Dobzhansky, T. Genetics of natural populations. XIV. A response of certain gene arrangements in the third chromosome of *Drosophila pseudoobscura* to natural selection. *Genetics* **32,** 142–160 (1947).
26. Saenko, S. V., Brakefield, P. M. & Beldade, P. Single locus affects embryonic segment polarity and multiple aspects of an adult evolutionary novelty. *BMC Biol.* **8,** 111 (2010).
27. van't Hof, A. E., Edmonds, N., Dalíková, M., Marec, F. & Saccheri, I. J. Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science* **332,** 958–960 (2011).
28. Kronforst, M. R. *et al.* Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proc. Natl Acad. Sci. USA* **103,** 6575–6580 (2006).
29. Chamberlain, N. L., Hill, R. I., Kapan, D. D., Gilbert, L. E. & Kronforst, M. R. Polymorphic butterfly reveals the missing link in ecological speciation. *Science* **326,** 847–850 (2009).
30. Merrill, R. M., Van Schooten, B., Scott, J. A. & Jiggins, C. D. Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. *Proc. R. Soc. B* **278,** 511–518 (2010).


**Supplementary Information** is linked to the online version of the paper


**Acknowledgements** We thank M. Blaxter and D. Charlesworth for advice throughout the study; The GenePool and S. Humphray for DNA sequencing; S. Kumar and A. Papanicolaou for bioinformatics support; M. Beltrán, A. Bulski, M. Veuille and the Botanique-Entomologie-Mycologie molecular facility (BoEM) for laboratory support; S. Johnston for genome-size estimates in *H. numata*; D. Obbard for providing R scripts; M. Abanto, S. Gallusser, C. Ramírez, L. de Silva, J. Barbut, B. Gilles and G. Lamas for help with butterfly rearing, fieldwork and collecting permits; and the Peruvian National Institute of Natural Resources (INRENA) for granting collecting and export permits (076-2007-INRENA-IFFS-DCB). Fieldwork in French Guiana was supported by a CNRS 'Nouragues Research Grant'. This work was supported by an EMBO long-term fellowship (ALTF-431-2004), EMBO-matching funds from NWO (Netherlands), a Royal Society University Research Fellowship (516002.K5917/ROG), a CNRS grant (ATIP Biodiversité 2008, France) and a European Research Council Starting Grant (ERC-Stg 'MimEvol') to M.J., a BBSRC grant (BBE0118451) to C.D.J. and R.H.ff.-C., a Leverhulme Trust grant (F/00144AY) to R.H.ff.-C., and a Royal Society University Research Fellowship and a Leverhulme Research Leadership grant to C.D.J.


**Author Contributions** M.J., C.D.J. and R.H.ff.-C. designed the study and contributed to all stages of the project. M.J., L. Frezal and R.T.J. performed the principal experiments and data analysis, with assistance from N.L.C., S.W.B., S.F.L., M.B., C.S., L. Ferguson, C.R.H., A.W. and P.A.W. BAC clone sequencing was carried out by C.D., R.G., C.L., R.C., H.B., S.S., J.R., M.C.J. and M.A.Q. M.J., A.W., C.D.J. and R.H.ff.-C. co-wrote the manuscript with input from all authors.

**Author Information** GenBank accessions for BAC clone sequences: FP885863, FP476061, FP565803, FP476023, CU856181, FP885878, FP476047, FP885857, CU856182, CU655868, FP885879, FP885861, FP885880, FP885855, CU914733, FP475989, CU655869, CU914734, CU633161, CU638865, CU856175, FP884220 and FP236755. Accessions for 1364 marker sequences: JN173798–JN175161.

## METHODS

**Chromosome walking.** A BAC library was constructed by Amplicon Express from five individual larvae from a polymorphic population of *H. numata* segregating for the forms *bicoloratus*, *tarapotensis* and *arcuella*. The BAC library has an estimated 119-kb average insert size and ×7 coverage of the 319-Mb genome. It was printed onto nylon filters and screened using 11 radiolabelled PCR probes designed to span the orthologous genomic sequences from an *H. melpomene* BAC library[9,11]. The BAC library was also fingerprinted by restriction digest, and overlapping clones were predicted from analysis of a fingerprinted contig (FPC) database. Fifty-three BAC clones were identified and tested for positive amplification of the original probes. Nineteen *H. numata* clones were sequenced by the Wellcome Trust Sanger Institute to high-throughput-genomic (HTG) phase 3 quality, totalling 2.9 Mb of overlapping genomic sequence. Because of the multiple chromosomal arrangements found in this region, and the fact that the clones sequenced come from chromosomes with differing gene orders as a consequence, it is impossible to reconstruct a single linear tilepath of clones at this stage. On the basis of reciprocal BLASTs of the clone sequences, the tilepath covers 0.95 Mb centred on *P*, consisting of four 'floating' contigs, two of which lie within the recombination interval of *P* and show different gene orders. A gap in the BAC chromosomal walk, estimated to be approximately 150 kb, remains between clones 7C9 and 14K13 despite extensive screening of the library. However, the gene markers *ARP-like* (*Hm00028*) and *Pros54* (*Hm00030*), predicted from the reference *Heliconius* sequence to lie in the middle of this gap, do indeed map in full linkage with other markers on either side of the gap, indicating the gene content of the gap is probably conserved. Two additional contigs of clones were identified and sequenced approximately 350 kb and 900 kb from the end of the tilepath further down the chromosome (Fig. 2a and Supplementary Table 2). Sequence similarity was plotted using the software Vista[31,32], with a 70% identity threshold and a 100-bp sliding window (Supplementary Fig. 2).

**Linkage mapping.** Male-informative markers were used to score crossing over events between markers linked and unlinked to *P*. Female Lepidoptera have achiasmatic meiosis, so crossing over only occurs during gametogenesis in the male parent. Thirty-six PCR-based markers were designed from BAC sequences and genotyped by visualization of differences in amplicon size, by restriction-fragment length polymorphism, or by sequence variation, in 366 individuals from six mapping families (B377, B465, B472, B502, B523 and A298 (refs 9, 11)). This was used to circumscribe the *P* supergene and orientate the tilepath by recombination. Special emphasis was given to the regions near the first crossover on either side of the supergene, for a precise positioning of the mapping boundary. The three recombinants with *P* on one side, and two recombinants on the other side, ensured that *P* was circumscribed and in full linkage with markers in an interval estimated to be about 400 kb (using a 0.92 scaling factor for *H. numata* versus *H. melpomene* genomic sequence, on the basis of sequences available for this region). The same procedure was used to score recombination events in unlinked regions using two sequenced BAC clones (bHN20L19 and bHM7E22) from unlinked chromosomes and three mapping families (Supplementary Table 1).

**BAC gene annotation.** Transcriptome sequencing, assembly and genomic annotation were performed as described elsewhere[11]. Briefly, annotation and gene predictions in all sequenced BACs (Supplementary Fig. 1 and Supplementary Table 3) were carried out with the Maker annotation pipeline[33], using 600K expressed sequenced tags generated with Roche 454 FLX pyrosequencing (200-base average read length) and assembled using the Mira assembler[34] to yield 24,992 objects. Complementary DNA was generated from wing-disc tissue extracted from juvenile stages from the Peruvian populations used throughout this study. Large variations in non-coding content, as well as in the presence and identity of transposable elements, were noted between clones within *H. numata* (Supplementary Table 4), causing some *H. numata* clones (for example, 14K13) to show a higher overall similarity to the orthologous *H. melpomene* sequences (bHM29B7) than to other orthologous *H. numata* clones (46M23 and 38G4). This indicates that the two *H. numata* chromosomal segments with differing gene orders may be anciently derived.

**Sequencing and breakpoint analysis.** Annotation of the BACs was used to identify exons within and outside the *P* mapping interval; exon markers were PCR-amplified and direct-sequenced for population-genetic analysis (Supplementary Table 8). Unlinked markers were chosen following the method in ref. 35. A total of 17 amplicons lying on linkage group 15, both within and outside the region containing *P*, plus 13 amplicons on unlinked chromosomes, were sequenced in 48–144 individuals (Supplementary Table 5) from a polymorphic population in eastern Peru. This populations segregates predominantly for the forms *silvana* and *aurora* (including sub-variants *isabellinus* and *elegans*). In subsequent analyses, we combined the *aurora* variants under the single class *aurora*, because the numerous phenotypic gradations that are found between these forms contrast with the clearly distinct phenotypic form *silvana*[4,15,36]. Furthermore, PCR assays of the rearrangement breakpoints and the absence of diagnostic nucleotide differences also indicated that these variants could be combined.

Breakpoint PCR assays were carried out with primers designed from exonic sequences on either side of each breakpoint (Supplementary Table 7). Fragments were amplified from 31 individuals with four different wing patterns (*bicoloratus*, *tarapotensis*, *aurora* and *silvana*) using long-range PCR (Qiagen) following the manufacturer's conditions, and were end-sequenced to confirm fragment identity. A second, larger sample of 201 individuals (including 161 from Peru and 40 from French Guiana) was assayed by standard (short-range) PCR amplification (Fermentas DreamTaq) using primers positioned closer to the breakpoints in unique non-coding DNA (Supplementary Table 7). Capillary sequencing of markers and breakpoint assays was performed on an ABI 3730 capillary sequencer with BigDye chemistry by the University of Edinburgh sequencing service (http://genepool.bio.ed.ac.uk/). Sequencing ambiguities were resolved manually using CodonCode Aligner (http://www.codoncode.com).

**Phenotype-by-genotype associations.** Association between genotype and phenotype was estimated for all polymorphic sites (SNPs) within each marker, taking into account the dominance relationships between different morphs. The Amazonian form of *H. numata* termed *silvana* (widespread from French Guiana to the Andean foothills) and its geographic replacement *illustris* (Andean valleys) are recessive to all other forms[4,9,36], which predicts that all *silvana* and *illustris* individuals are homozygous at *P*, whereas individuals of other, co-occurring forms can be either heterozygous or homozygous for a different allele. We estimated the association between genotype and phenotype by testing the hypothesis that the major allele should be homozygous in *silvana* individuals and heterozygous or absent in *aurora* individuals. For each polymorphic site with a minor allele frequency of at least 0.1, we calculated the proportion of individuals whose genotype conformed to this hypothesis and tested the null hypothesis of no phenotype–genotype association using Fisher's exact test. This method allows testing of the phenotype-by-genotype association while using knowledge of the dominance relationships between *P* alleles.

**Linkage disequilibrium (LD) across the *P* supergene.** LD was assessed for all pairs of polymorphic sites within and surrounding the *P* region, as well as in 12 unlinked markers (Fig. 2c) in 59 specimens from a polymorphic population near Yurimaguas, eastern Peru (Supplementary Table 5). We used genotypic correlation-based testing for LD with several alleles with unknown phase between sites[37]. Significance was tested by a permutation test for the genotypic correlation statistic $r^2$ (ref. 37).

**Haplotype networks.** The phase of SNP variation was determined by analysing the segregation of alleles from parent to offspring in markers sequenced from the mapping families, which originate from the same populations as the population samples. Haplotypes from the population samples were inferred by coalescent-based Bayesian methods using the PHASE 2.1 algorithm[38], optimized by including sequences with known phase from mapping families. The high level of LD across the region ensured a robust inference of haplotypic diversity for all markers. Haplotype networks were constructed by parsimony using the Network package[39,40] (http://www.fluxus-engineering.com/). The level of genetic differentiation between the *silvana* and non-*silvana* groups was estimated in DNAsp[41] using the Fst statistics[42], and its significance was tested by permutation (Supplementary Fig. 5). Fst was not used to test for genetic differentiation among different populations, but rather to assess genetic differentiation between two groups of individuals in a single population (*silvana* versus non-*silvana*). Even high levels of genetic differentiation do not indicate the absence of random mating, but rather that these sites are linked and/or that they co-vary with the functional loci that determine the different mimetic morphs.


31. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32,** W273–W279 (2004).
32. Mayor, C. *et al.* VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046 (2000).
33. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18,** 188–196 (2008).
34. Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14,** 1147–1159 (2004).
35. Salazar, C., Jiggins, C. D., Taylor, J. E., Kronforst, M. R. & Linares, M. Gene flow and the genealogical history of *Heliconius heurippa*. *BMC Evol. Biol.* **8,** 132 (2008).
36. Brown, K. S. An illustrated key to the silvaniform *Heliconius* (Lepidoptera: Nymphalidae) with descriptions of new subspecies. *Trans. Am. Entomol. Soc.* **102,** 373–484 (1976).
37. Zaykin, D. V., Pudovkin, A. & Weir, B. S. Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics* **180,** 533–545 (2008).
38. Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73,** 1162–1169 (2003).

39. Polzin, T. & Daneshmand, S. V. On Steiner trees and minimum spanning trees in hypergraphs. *Oper. Res. Lett.* **31,** 12–20 (2003).
40. Forster, P., Torroni, A., Renfrew, C. & Rohl, A. Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol. Biol. Evol.* **18,** 1864–1881 (2001).
41. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25,** 1451–1452 (2009).
42. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA-sequence data. *Genetics* **132,** 583–589 (1992).