LETTER

doi:10.1038/nature13112

doublesex is a mimicry supergene

K. Kunte[1]*, W. Zhang[2]*, A. Tenger-Trolander[2], D. H. Palmer[3], A. Martin[4], R. D. Reed[4], S. P. Mullen[5] & M. R. Kronforst[2,3]

One of the most striking examples of sexual dimorphism is sex-limited mimicry in butterflies, a phenomenon in which one sex—usually the female—mimics a toxic model species, whereas the other sex displays a different wing pattern[1]. Sex-limited mimicry is phylogenetically widespread in the swallowtail butterfly genus *Papilio*, in which it is often associated with female mimetic polymorphism[1–3]. In multiple polymorphic species, the entire wing pattern phenotype is controlled by a single Mendelian 'supergene'[4]. Although theoretical work has explored the evolutionary dynamics of supergene mimicry[5–9], there are almost no empirical data that address the critical issue of what a mimicry supergene actually is at a functional level. Using an integrative approach combining genetic and association mapping, transcriptome and genome sequencing, and gene expression analyses, we show that a single gene, *doublesex*, controls supergene mimicry in *Papilio polytes*. This is in contrast to the long-held view that supergenes are likely to be controlled by a tightly linked cluster of loci[4]. Analysis of gene expression and DNA sequence variation indicates that isoform expression differences contribute to the functional differences between *dsx* mimicry alleles, and protein sequence evolution may also have a role. Our results combine elements from different hypotheses for the identity of supergenes, showing that a single gene can switch the entire wing pattern among mimicry phenotypes but may require multiple, tightly linked mutations to do so.

Wing pattern mimicry in butterflies, a phenomenon in which natural selection by predators causes unrelated species to evolve similar wing patterns, has served as an important model for studying adaptation since the earliest days of modern evolutionary theory[10]. Classical Batesian mimicry, in which an undefended mimic evolves to look like a toxic model, is a parasitic relationship in which the mimic gains an advantage at the expense of the model. Such systems have well-characterized frequency dependence[1,7], sometimes resulting in sexual dimorphism and mimetic polymorphism[1–3,8,11,12]. Swallowtail butterflies in the genus *Papilio* are well-known Batesian mimics, providing some of the most extreme examples of sexual dimorphism and polymorphism among living organisms[1,2,12]. For instance, in the species *Papilio polytes*, males all display the same non-mimetic wing pattern, whereas females display either a male-like pattern (form *cyrus*) or one of several different patterns that mimic toxic species in the genus *Pachliopta* (Fig. 1). Female wing pattern is polymorphic in local areas and there are no intermediate forms. The early crossing experiments of Clarke and Sheppard[13] revealed that variation in the entire wing pattern, as well as the presence versus absence of hindwing 'tails', is controlled by a single Mendelian locus, with female polymorphism resulting from multiple alleles, each with its place in a dominance hierarchy. Clarke and Sheppard also showed that the mimicry locus is autosomal, so sexual dimorphism is not directly mediated by sex linkage in this case[13].

This phenomenon, in which the entire wing pattern is controlled by a single Mendelian locus, is referred to as 'supergene' mimicry[4]. Because Clarke and Sheppard occasionally witnessed individuals with putatively recombinant wing patterns, they envisioned a supergene as a tightly linked cluster of loci, each controlling a distinct subset of the wing pattern. However, Clarke and Sheppard found virtually no evidence for recombination in *P. polytes*[13], although they did recover apparently recombinant phenotypes in other species, such as *P. memnon*[14]. Over the past few decades, supergene mimicry has received considerable theoretical attention[5–8], but there are almost no empirical data that address the molecular basis of a supergene. One example from *Heliconius* butterflies, which involves supergene mimicry but not sexual dimorphism, suggests that supergenes may be the result of chromosomal inversions that lock multiple adjacent genes into a single, non-recombining unit[15].

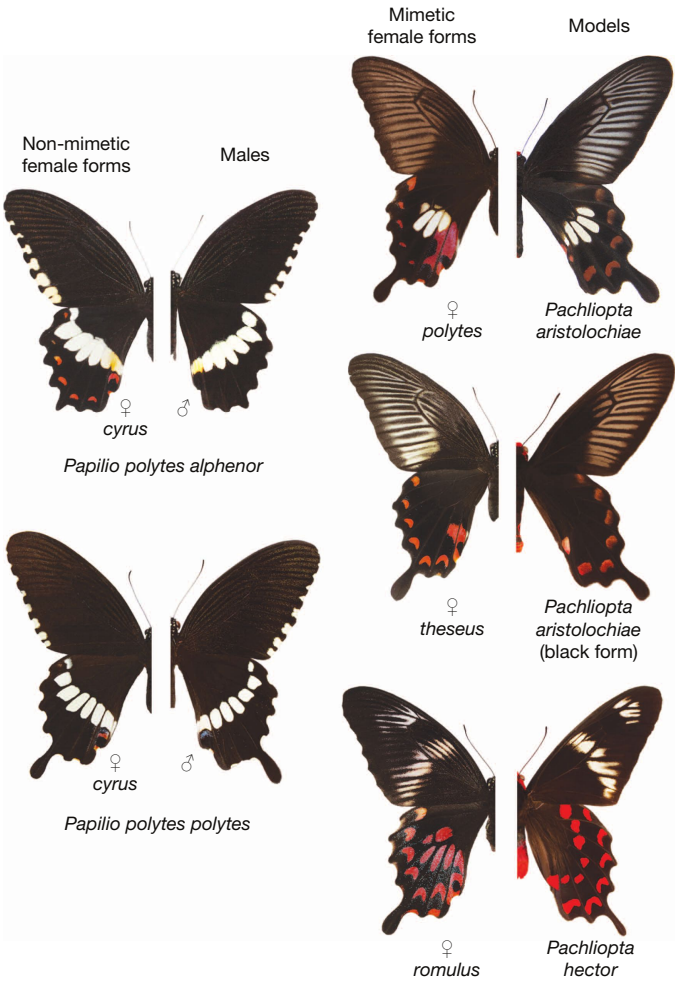

Figure 1 | Polymorphic, sex-limited mimicry in *Papilio polytes*. Non-mimetic (form *cyrus*) females look like males, whereas mimetic female morphs (forms *polytes*, *theseus* and *romulus*) mimic distantly related, toxic *Pachliopta* swallowtails. The presence of hindwing tails on males and *cyrus* females is variable among *P. polytes* populations. Our analyses focused on *P. polytes alphenor*, a group lacking tails on non-mimetic butterflies, and presence versus absence of tails segregated perfectly with female wing pattern in our crosses.

[1]National Center for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru 560065, India. [2]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [3]Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois 60637, USA. [4]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [5]Department of Biology, Boston University, Boston, Massachusetts 02215, USA.
*These authors contributed equally to this work.

Other evolutionary phenomena that involve supergene-like genetic architectures, such as self-incompatibility in plants and segregation distortion in *Drosophila melanogaster*, have also been traced back to multiple linked genes[16,17]. A second possibility is that a master regulator could gain control of the distinct networks that pattern various aspects of the wing, and hence control the entire phenotype from a single locus. Although this single-gene hypothesis has been discussed[6,9,18], there are no empirical data to support it.

Using a multi-step genetic mapping process that involved rearing nine F_2 backcross families (Fig. 2a), bulk segregant analysis with restriction-site associated DNA (RAD) markers, screening and sequencing bacterial artificial chromosome (BAC) clones, and fine-mapping, we mapped the mimicry locus in *P. polytes* back to a 300-kilobase (kb) region of the genome that contained five genes (Fig. 2b). We were intrigued to find that one of these genes was *doublesex* (*dsx*), a transcription factor in insects that controls somatic sex differentiation by alternative splicing[19,20]. In *Drosophila*, *dsx* is alternatively spliced into two isoforms: a male-specific form that leads to male sexual differentiation, and an alternative female form that causes female sexual differentiation[19–21]. In other insects, *dsx* functions the same way although there can be more than one male and female isoform[22,23].

On the basis of our mapping data and the known role of *dsx* in mediating sexual dimorphism[23–26], we proposed that *dsx* might control both the sex-limited and female polymorphism components of *P. polytes* mimicry. To test this hypothesis, we generated a reference genome sequence across our target interval and performed comprehensive association mapping by re-sequencing the genomes of 15 mimetic (form *polytes*) and 15 non-mimetic (form *cyrus*) butterflies (Extended Data Table 1). This yielded multiple perfect associations in *dsx* but only weak associations immediately outside of *dsx* (Fig. 2c). A separate genome-wide association study (GWAS) also yielded *dsx* as the top association hit (Extended Data Table 2). Long-term balancing selection, which maintains mimicry polymorphism[1,8,12], is expected to result in a localized excess of nucleotide variation driven by the accumulation of neutral substitutions on alternative alleles[27]. Analysis of DNA sequence variation revealed a highly significant excess of nucleotide polymorphism in *dsx*, relative to neighbouring genes (Table 1 and Extended Data Table 3), and comparisons between mimetic and non-mimetic individuals revealed over 1,000 nucleotide substitutions differentiating mimetic and non-mimetic *dsx* alleles (Table 1). This is in contrast to all neighbouring genes, which show little polymorphism and no fixed differences between mimicry forms.

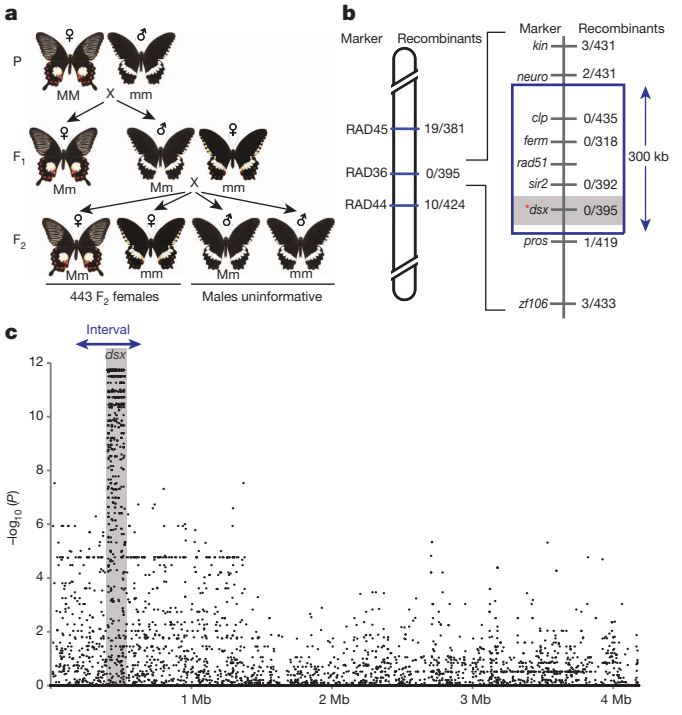

Figure 2 | Mapping the mimicry supergene. **a**, A series of nine backcross families yielded a total of 443 F_2 females that segregated 1:1 for female mimicry phenotype. **b**, Genome-wide mapping with RAD markers and subsequent fine-mapping localized the mimicry locus to a 300-kb interval containing five genes, one of which was *doublesex* (*dsx*). **c**, Association mapping, based on full genome sequences of 30 *P. polytes* butterflies, revealed multiple perfect associations inside *dsx* but none outside the gene. The positions of the 300-kb zero-recombinant interval and *dsx* are indicated. Data points represent false-discovery rate (FDR)-adjusted *P* values for a total of 94,776 SNPs.

The involvement of *dsx* as the mimicry supergene indicates a potential role for alternative splicing in the control of wing pattern. Transcriptome assembly based on wing-disc-derived RNA yielded three distinct female *dsx* isoforms and one male isoform (Fig. 3a). However, cloning and sequencing *dsx* isoforms from mimetic and non-mimetic males and females yielded the same repertoire of isoforms in butterflies with alternative mimicry alleles. Comparisons of isoform expression using quantitative reverse transcription PCR (qRT–PCR) revealed that all three female isoforms show strong female-biased expression (Fig. 3b–d). Two of these, isoforms 1 and 2, further showed pronounced wing-biased expression, whereas the third female isoform had body-biased expression (Fig. 3b–d). Comparisons between mimetic and non-mimetic females for wing-biased isoforms 1 and 2 revealed marked upregulation in mimetic females relative to non-mimetic females (Fig. 3e, f). This biased expression probably contributes to the functional difference between mimicry alleles. Notably, expression of isoforms 1 and 2 seems to increase at day 5 after pupation (Fig. 3g), a stage at which immunodetection of Dsx spatial expression on mimetic forewings revealed a marked spatial correspondence with adult wing pattern (Fig. 3h).

Overall, our results indicate a surprising mode of action for *dsx* as a mimicry supergene. As a classic example of alternative splicing, our initial hypothesis was that alternative splicing would also underlie the phenotypic switch between female wing patterns. Although we do find clear evidence of alternative splicing, and different levels of isoform expression between female wing patterns, the set of female isoforms does not differ between groups. Rather, gene expression variation seems to have a central role in controlling mimicry polymorphism. Another striking feature of *dsx* in *P. polytes* is the large number of nucleotide substitutions that differ between mimicry alleles. The accumulation of neutral substitutions that is expected from balancing selection makes it difficult to infer which of these changes might be functionally related to mimicry polymorphism. However, we note that the proportion of fixed differences between *cyrus* and *polytes* haplotypes is over seven times greater in coding regions (72 out of 1,068 differences) compared to non-coding regions (972 out of 108,036 differences), and these coding region changes include 25 amino acid substitutions located primarily in the first exon (Table 1). The amino acid changes in exon 1 are clustered in two regions: the 5′ end of the protein, in front of the DNA binding (DM) domain, and the region between the DM domain and the dimerization domain; there are no amino acid changes in either domain (Extended Data Fig. 1). To explore the potential impact of these amino acid substitutions, we predicted secondary and tertiary structures for both the *cyrus* and *polytes* Dsx proteins and found that they differ markedly—the non-mimetic *cyrus* protein folds much like other insects, such as *Bombyx mori*, whereas the mimetic *polytes* protein structure is highly divergent (Extended Data Fig. 2). In addition to the differential expression of female isoforms, we speculate that distinct Dsx protein structures may also contribute to female polymorphism, with alternative alleles differentially regulating different downstream targets as a result of divergent DNA or coactivator binding properties.

How are a large number of nucleotide substitutions maintained in complete linkage disequilibrium over the approximately 100-kb length of *dsx*? Recombination between mimicry alleles in heterozygotes should break up *dsx* haplotypes, and the fact that we see many differences

Table 1 | DNA sequence variation in *Papilio polytes* near the mimicry supergene

Gene	Section	Length (bp)	Fixed synonymous/silent substitutions	Fixed non-synonymous substitutions	Total SNPs
neuro	ORF	519	0	0	8
clp	ORF	1,443	0	0	26
ferm	ORF	2,181	0	0	17
rad51	ORF	1,017	0	0	2
sir2	ORF	1,224	0	0	3
dsx	Exon 1	588	31	21	59
	Exon 2	144	5	0	8
	Exon 3	84	0	0	1
	Exon 4	69	2	1	3
	Exon 5	183	9	3	13
	Non-coding	108,036	972	NA	6,781
pros	ORF	2,895	0	0	21

Counts of synonymous/silent and non-synonymous nucleotide substitutions fixed between mimetic (*polytes*) and non-mimetic (*cyrus*) *P. polytes* butterflies in genes located near the mimicry supergene, as well as the total number of SNPs in each gene. Counts for *dsx* are separated by gene section (exons, non-coding) whereas counts for other genes represent predicted open reading frame (ORF).

between mimicry alleles suggests that something is reducing recombination immediately around *dsx*. Chromosomal inversions are well known to reduce recombination in heterozygotes[28], making this a likely explanation. We first verified that the *dsx* region does indeed exhibit elevated linkage disequilibrium relative to adjacent regions (Extended Data Fig. 3), and then we searched for evidence of structural variation around *dsx* using our genome re-sequencing data. As predicted, we found support for an inversion polymorphism associated with mimicry alleles, the breakpoints of which flank *dsx* (Extended Data Table 4 and Extended Data Fig. 4). Given the long history of speculation about the molecular identity of supergenes, it is interesting that we have uncovered a scenario that unites both possible explanations: reduced recombination among presumably different functional elements and single gene control. In essence, the multiple, tightly linked loci proposed by Clarke and Sheppard[13] may, in this case, actually be multiple, tightly linked mutations in the same gene.

It is perhaps unexpected that a gene so intimately connected to an essential developmental process could be co-opted to also control intraspecific polymorphism. Somehow, *dsx* has retained its highly conserved sex-differentiation properties[19–21] while also evolving new phenotype-switching properties in just one sex. Our results suggest two complementary mechanisms that may underlie the ability of *dsx* to have two distinct roles in *P. polytes*. First, although we found many mutations in the Dsx protein, none of these occurs in the DM or dimerization domains, which are essential components for its ancestral function in sexual differentiation. Second, we also found that different *dsx* isoforms are expressed on the wings and in the body of females, which may also allow this one gene to carry out a novel function on the wings.

R. A. Fisher called mimicry the "greatest post-Darwinian application of Natural Selection"[7] and supergene mimicry stands out as a particularly extreme adaptive endpoint. Although little is known about the molecular and developmental basis of supergene mimicry, previous evidence suggests that multiple, tightly linked genes probably underlie this phenomenon. Here we have integrated multiple approaches to reveal that a single gene acts as the mimicry supergene in *P. polytes*. In so doing, we have greatly expanded the known role of *doublesex* and the sexual differentiation pathway generally. Female-limited mimetic polymorphism has evolved independently multiple times in the genus *Papilio*[2], making this a useful system in which to investigate the generality of our results. One might predict that the sex determination pathway, and *dsx* in particular, may have been co-opted repeatedly to control this phenomenon because this pathway is preconfigured to mediate the most widespread polymorphism in the animal kingdom—sex. Interestingly, available data, although limited, suggest that this is not the case. For instance, female mimetic polymorphism in *Papilio dardanus* has previously been mapped to a genomic region containing the genes *engrailed* and *invected*[29], which is not linked to *dsx*. Furthermore, female mimetic

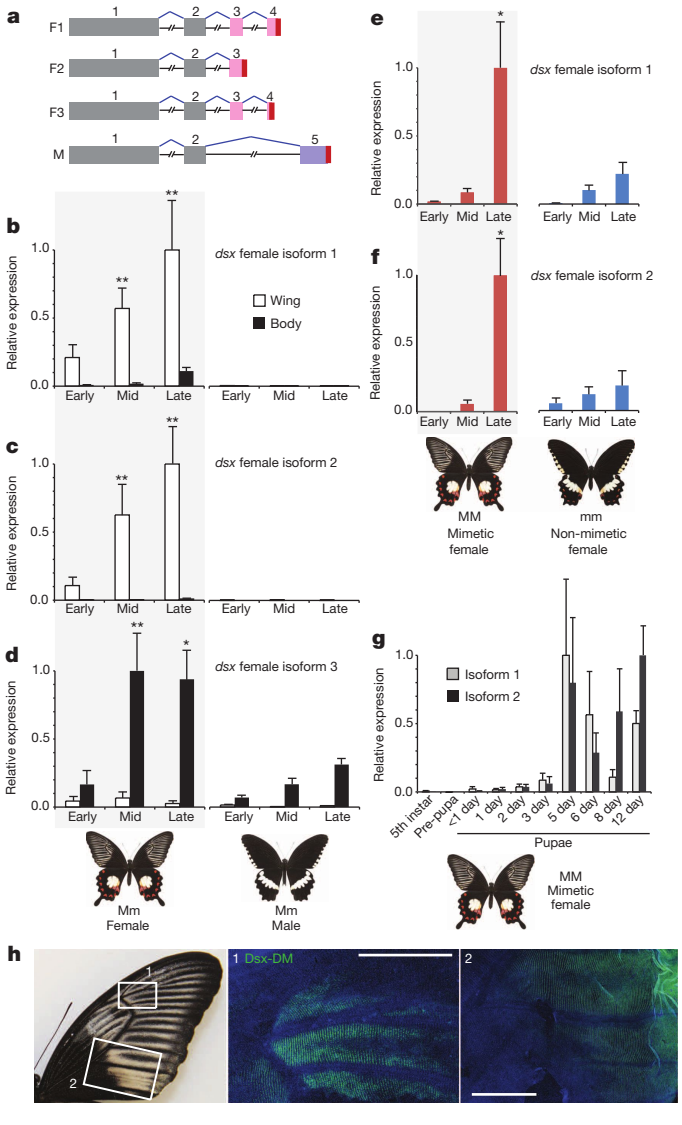

Figure 3 | Expression of *doublesex* in *P. polytes*. **a**, *dsx* is alternatively spliced into three female isoforms and one male isoform. **b**–**d**, Expression of female isoforms is strongly female-biased and isoform 1 (**b**) and isoform 2 (**c**) show wing-biased expression whereas isoform 3 expression (**d**) is body-biased; $n = 6$ (female early), 6 (female mid), 5 (female late), 6 (male early), 6 (male mid), 7 (male late). **e**, **f**, Female isoforms 1 and 2 also show elevated expression in mimetic females (*polytes*) relative to non-mimetic females (*cyrus*); $n = 9$ (*polytes* early), 9 (*polytes* mid), 12 (*polytes* late), 3 (*cyrus* early), 3 (*cyrus* mid), 3 (*cyrus* late). **g**, Finer scale temporal data for isoforms 1 and 2 on mimetic female wings suggests expression of both increases at 5 days after pupation; $n = 3$ for each time point. **h**, Immunodetection of Dsx protein on mimetic female wings 5 days after pupation reveals strong correlation with adult wing pattern. Scale bars, 1 mm. Data represented as mean ± s.e.m. All *n* values indicate number of biological replicates. $^*P < 0.05$; $^{**}P < 0.01$, ANOVA and Tukey's HSD test.

polymorphism in *Papilio glaucus* is sex-linked, with the primary switch locus on the W chromosome and a modifier on the Z chromosome[30]. Future work will determine whether other instances of sex-limited polymorphism, in butterflies and beyond, involve the sex differentiation pathway, but evolution, it seems, can take many paths to the extreme supergene genetic architecture, even among members of the same genus.

METHODS SUMMARY

Using one backcross mapping family (94 females: 48 *cyrus* and 46 *polytes*), we performed bulk segregant analysis with RAD markers. Subsequent fine mapping, using a total of nine mapping families (443 females: 229 *cyrus* and 214 *polytes*), and BAC sequencing isolated the mimicry locus to a 300-kb interval containing five genes, one of which was *dsx*. We sequenced the genomes of 30 laboratory-reared individuals (15 *polytes* and 15 *cyrus*) with an Illumina HiSeq 2000 and generated a reference genome sequence for *P. polytes* using both *de novo* and reference-guided assembly. Single nucleotide polymorphism (SNP) calling of the 30 sequenced genomes yielded 675,526 genome-wide SNPs and 94,776 SNPs across a 4-megabase (Mb) scaffold containing *dsx*. GWAS was performed by calculating genetic differentiation (F_{ST}) between *polytes* and *cyrus* individuals for *de novo* assembly scaffolds. Association tests across the 4-Mb *dsx* scaffold were performed using a false-discovery rate correction. We used Hudson–Kreitman–Aguadé (HKA) tests to compare nucleotide polymorphism among genes in the mimicry supergene region. Pairwise linkage disequilibrium was calculated among biallelic SNPs in two different portions of the *dsx* scaffold, and we used the short read sequence data to perform structural variant detection. We then used BLAST to identify scaffolds from a *de novo* assembly of *polytes* samples that appear to span an inversion containing *dsx*. Subsequent PCR tests isolated the 3′ breakpoint to a 2-kb interval. RNA-seq data, generated from wing-disc-derived *P. polytes* RNA, were used to perform transcriptome assembly and qRT–PCR was used to measure *dsx* isoform expression in males and females across development. We used a protein homology web server to infer secondary and tertiary structures of *polytes* and *cyrus* Dsx proteins, as well Dsx from *Bombyx mori*. Immunodetection of Dsx was carried out using a monoclonal anti-*Drosophila* Dsx DM domain antibody.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 November 2013; accepted 30 January 2014.
Published online 5 March 2014.

1. Joron, M. & Mallet, J. L. Diversity in mimicry: paradox or paradigm? *Trends Ecol. Evol.* **13,** 461–466 (1998).
2. Kunte, K. The diversity and evolution of batesian mimicry in *Papilio* swallowtail butterflies. *Evolution* **63,** 2707–2716 (2009).
3. Kunte, K. Female-limited mimetic polymorphism: a review of theories and a critique of sexual selection as balancing selection. *Anim. Behav.* **78,** 1029–1036 (2009).
4. Clarke, C. A. & Sheppard, P. M. Super-genes and mimicry. *Heredity* **14,** 175–185 (1960).
5. Charlesworth, D. & Charlesworth, B. Theoretical genetics of Batesian mimicry II. Evolution of supergenes. *J. Theor. Biol.* **55,** 305–324 (1975).
6. Charlesworth, D. & Charlesworth, B. Mimicry: the hunting of the supergene. *Curr. Biol.* **21,** R846–R848 (2011).
7. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press, 1930).
8. Sheppard, P. M. The evolution of mimicry: a problem in ecology and genetics. *Cold Spring Harb. Symp. Quant. Biol.* **24,** 131–140 (1959).
9. Turner, J. R. G. in *The Biology of Butterflies* (eds Vane-Wright, R. I. & Ackery, P. R.) 141–161 (Academic, 1984).
10. Bates, H. W. Contributions to an insect fauna of the Amazon valley (Lepidoptera: Heliconidae). *Trans. Linn. Soc. (Lond.)* **23,** 495–566 (1862).
11. Ford, E. B. The genetics of polymorphism in the Lepidoptera. *Adv. Genet.* **5,** 43–87 (1953).
12. Mallet, J. & Joron, M. Evolution of diversity in warning color and mimicry: Polymorphisms, shifting balance, and speciation. *Annu. Rev. Ecol. Syst.* **30,** 201–233 (1999).
13. Clarke, C. A. & Sheppard, P. M. The genetics of the mimetic butterfly *Papilio polytes* L. *Phil. Trans. R. Soc. Lond. B* **263,** 431–458 (1972).
14. Clarke, C. A., Sheppard, P. M. & Thornton, I. W. B. The genetics of the mimetic butterfly *Papilio memnon* L. *Phil. Trans. R. Soc. Lond. B* **254,** 37–89 (1968).
15. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477,** 203–206 (2011).
16. Larracuente, A. M. & Presgraves, D. C. The selfish *Segregation Distorter* gene complex of *Drosophila melanogaster*. *Genetics* **192,** 33–53 (2012).
17. Takayama, S. & Isogai, A. Self-incompatibility in plants. *Annu. Rev. Plant Biol.* **56,** 467–489 (2005).
18. Nijhout, H. F. Developmental perspectives on evolution of butterfly mimicry. *Bioscience* **44,** 148–157 (1994).
19. Burtis, K. C. & Baker, B. S. *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell* **56,** 997–1010 (1989).
20. Williams, T. M. & Carroll, S. B. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nature Rev. Genet.* **10,** 797–804 (2009).
21. Kopp, A. *Dmrt* genes in the development and evolution of sexual dimorphism. *Trends Genet.* **28,** 175–184 (2012).
22. Cho, S., Huang, Z. Y. & Zhang, J. Z. Sex-specific splicing of the honeybee *doublesex* gene reveals 300 million years of evolution at the bottom of the insect sex-determination pathway. *Genetics* **177,** 1733–1741 (2007).
23. Kijimoto, T., Moczek, A. P. & Andrews, J. Diversification of *doublesex* function underlies morph-, sex-, and species-specific development of beetle horns. *Proc. Natl Acad. Sci. USA* **109,** 20526–20531 (2012).
24. Tanaka, K., Barmina, O., Sanders, L. E., Arbeitman, M. N. & Kopp, A. Evolution of sex-specific traits through changes in HOX-dependent *doublesex* expression. *PLoS Biol.* **9,** e1001131 (2011).
25. Williams, T. M. *et al.* The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *Drosophila*. *Cell* **134,** 610–623 (2008).
26. Loehlin, D. W. *et al.* Non-coding changes cause sex-specific wing size differences between closely related species of *Nasonia*. *PLoS Genet.* **6,** e1000821 (2010).
27. Charlesworth, B. & Charlesworth, D. *Elements of Evolutionary Genetics* (Roberts & Co., 2010).
28. Hoffmann, A. A., Sgro, C. M. & Weeks, A. R. Chromosomal inversion polymorphisms and adaptation. *Trends Ecol. Evol.* **19,** 482–488 (2004).
29. 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. Lond. B* **275,** 1181–1188 (2008).
30. Scriber, J. M., Hagen, R. H. & Lederhouse, R. C. Genetics of mimicry in the tiger swallowtail butterflies, *Papilio glaucus* and P. *canadensis* (Lepidoptera: Papilionidae). *Evolution* **50,** 222–236 (1996).

Acknowledgements We thank W. Wang for sharing genome sequence data, C. Robinett for providing the Dsx-DM monoclonal antibody, and E. Westerman, S. Nallu, M. Zhang, G. Garcia and N. Pierce for assistance and discussion. This project was funded by National Science Foundation grant DEB-1316037 to M.R.K.

Author Contributions K.K. conceived the project and helped design the study, reared mapping families and samples for gene expression analysis and genome sequencing, performed bulk-segregant analysis and RAD mapping, and contributed to drafting the manuscript. W.Z. generated the reference genome sequences and transcriptome assemblies, performed association mapping, GWAS analysis, HKA tests, structural variant detection and linkage disequilibrium analyses, analysis of protein structure and synonymous/non-synonymous calculations, and contributed to drafting the manuscript. A.T.-T. assisted with butterfly husbandry, performed fine mapping, cDNA sequencing and qRT–PCR analyses. D.H.P. performed qRT–PCR analyses. A.M. and R.D.R. performed Dsx immunohistochemistry. S.P.M. helped design the project and contributed to drafting the manuscript. M.R.K. designed and directed the project, analysed data and wrote the manuscript.

Author Information Sequence data are available from NCBI SRA (SRP035394) and GenBank (KJ150616–KJ150623). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R.K. (mkronforst@uchicago.edu) or K.K. (krushnamegh@ncbs.res.in).

METHODS

Butterfly husbandry. Our polymorphic *P. polytes* laboratory colony was initially founded with individuals supplied by Flora Farm Butterfly in the Philippines. Butterflies were maintained as large, outbred colonies, founded using a large number of wild-caught adults. Across most of its range, male and non-mimetic female *P. polytes* have hindwing 'tails' like the mimetic forms. However, on some islands in the Philippines, males and non-mimetic females have lost tails whereas they are retained on mimetic females (because the model species they are mimicking have tails). Our starting laboratory colony exhibited this tail polymorphism, which segregates with wing pattern.

Starting from our initial laboratory colony, we generated a pure mimetic (form *polytes*) and a pure non-mimetic (form *cyrus*) line by identifying and grouping individuals of either homozygous genotype from multiple independent crosses. For genetic mapping, we generated mimicry heterozygous F_1 offspring by crossing males and females from our pure lines. F_1 heterozygous males were subsequently backcrossed to females from the pure *cyrus* line to generate broods segregating 1:1 for mimicry phenotype in females (males all display the same non-mimetic phenotype regardless of mimicry genotype). We used F_1 males only because female butterflies undergo achiasmatic oogenesis which results in no recombination between homologous chromosomes during meiosis. Recombination does occur during meiosis in males, which facilitates linkage mapping at finer scales. In total, we generated nine backcross mapping families, yielding a total of 443 backcross females, 229 *cyrus* and 214 *polytes*. Because all backcross males exhibit the same phenotype, regardless of genotype, we did not use males for genetic mapping of the mimicry locus.

Bulk segregant analysis. Mapping family 1 was a cross between a mimicry heterozygous male (PP555) and a homozygous recessive (*cyrus*) female (PP566). This cross yielded 94 female offspring, 48 *cyrus* and 46 *polytes*. We pooled DNA, in equimolar amounts, of the female offspring by wing pattern and generated RAD tag data for four resulting samples: parent PP555, parent PP566, *cyrus* female pool, *polytes* female pool. RAD-seq yielded approximately 340,000 raw reads for each parent and approximately 1.75 million reads for each offspring pool. Subsequent data analysis resulted in 3,515 markers that were heterozygous in PP555 and homozygous in PP566, 38 (1.08%) of which were significantly ($P < 0.01$, Fisher's exact test) associated with phenotype based on read counts in the offspring pools. Another 1,171 markers appeared to be fixed between PP555 and PP566, of which 9 (0.77%) were associated with wing pattern. We assembled contigs, using the Illumina paired-end data, for nine RAD markers most strongly associated with mimicry. We then designed PCR primers for each RAD, PCR amplified and Sanger sequenced each in all backcross parents (RAD7F, AGGTGWTATACGCGTGATCTAAACACG; RAD7R, GATCTCTGCTTTAGAATTAATCG; RAD15F, ATACCGTCCACGCGGAATTG; RAD15R, ACCGGAGCTGCTCTCAAACTACC; RAD19F, GCCACCTGCACCGCCTCCGCG; RAD19R, TTACTTTAGTGCGTACTTACTTACG; RAD31F, TCTCCTTACGTTTAATTGACTAC; RAD31R, CGAAGTCGCGAGCAACAACTAG; RAD36F, GCGAAATTGTTTCGAAAATAG; RAD36R, CTTTATTGTGTTTTTTACCCGGCTC; RAD40F, GGCCCTTACAAKTGTTAATTG; RAD40R, CAGRACACTAAAAAGTAAC; RAD43F, GTCGACCGTGGTGGCTTTCTAATGTCG; RAD43R, ATTACTATTATTCACAGATAAGC; RAD44F, CATATAGTATTCATCGACTTG; RAD44R, CTCTTACACCGTCAAATCCACGTTC; RAD45F, CTATGYGTTGTTAAGGACTTACG; RAD45R, TGGTCGTGGTATTACACCGGGCGTAG). We ultimately genotyped three markers in all 443 backcross females. RAD36 was successfully genotyped in 395 offspring, RAD44 in 424 offspring, and RAD45 in 381 offspring. RAD36 was found to perfectly co-segregate with mimicry, whereas RAD44 and RAD45 flanked the mimicry locus with 10 out of 424 recombinants identified at RAD44 and 19 out of 381 recombinants at RAD45. Because a unique set of individuals was recombinant at RAD44 and RAD45, they appear to sit on opposite sides of the mimicry locus.

BAC library screening. We generated a BAC library for *P. polytes* with the company Amplicon Express. The library consisted of 36,864 clones with an average insert size of 115 kb (15× genome coverage). We screened this library with probes designed from the three mapped RAD markers, which yielded 19 clones positive for RAD36, 12 clones positive for RAD44, and 11 clones positive for RAD45. By end-sequencing and BLASTing against the protein nr database, we recovered portions of 16 putative protein-coding genes. One of these was *doublesex*, which we found in an end sequence of clone PPOL20B02, identified with the RAD36 probe. In addition to *dsx*, we identified five putative protein-coding genes in BAC end sequences that mapped uniquely to the genome sequence of *Bombyx mori*; these and *dsx* are located on *B. mori* scaffold nscaf2823 on chromosome 25. We sequenced one BAC from each of the three probes in its entirety (PPOL2F2, PPOL8G9, 4B14) using short-read Illumina sequencing and these were used to verify genome location. All three BAC sequences BLASTed to *B. mori* scaffold nscaf2823. We subsequently screened the BAC library with probes designed from markers used for fine-mapping (see below: *clp*, *ferm*, *sir2*, *neuro*, *pros*) and identified 96 clones positive for one or more marker. All 96 clones were sequenced at low coverage by Amplicon Express.

Fine mapping. Using the *B. mori* genome sequence, we identified seven genes surrounding *doublesex*: *kinesin KIF4* (*kin*), *Neuroendocrine convertase 2* (*neuro*), *ATP-dependent Clp protease proteolytic subunit* (*clp*), *Fermitin 1* (*ferm*), *NAD-dependent deacetylase sirtuin 2* (*sir2*), *prospero* (*pros*), and *zinc finger protein 106* (*zf106*). Using transcriptome data from *P. polytes* (see below) we designed PCR primers for all seven genes (KinF ATATCTGATCTGAAGAAGAA, KinR TCATTGCGAAGACGCAGCAT; NeuroF CGTTTCCACTGGACTATGAA, NeuroR GTGCTCCAGGTACCGCACCT; ClpF AACTGGACGAGGTGAGAGA, ClpR TCAATCAATCCAAAGGCTTT; FermF GCCTGTCGCTTCCAATCACA, FermR GACTCCCTGGACTGAGAGTC; Sir2F TGGAACCTCTCTGGCAAACT, Sir2R GGCTAAAACAACGAAATCTACG; ProsF GGACACGAATCGGAGACTGT, ProsR GCCTCTGTTGCTGGCTATTC; zf106F CAAGAATGAAGGAAATAGAT, zf106R TCGTCTAATGTTATTATTTC). We PCR amplified and Sanger sequenced each gene fragment in all mapping families and used SNPs to fine map the mimicry locus. The final zero-recombinant interval contained five genes: *clp*, *ferm*, *sir2*, *dsx* and a fifth gene which we did not use as a marker for fine mapping, the DNA repair protein *RAD51*.

Genome re-sequencing. We extracted genomic DNA from 30 laboratory-reared individuals (15 *polytes* and 15 *cyrus*). Illumina paired-end libraries were constructed using the Illumina TruSeq protocol and sequenced with an Illumina HiSeq 2000. Low-quality data were filtered from raw reads and only high-quality reads were used for downstream analyses (Extended Data Table 1).

GWAS. We performed *de novo* genome assembly using SOAPdenovo2 pregraph_sparse_63mer v1.0.3 and SOAPdenovo-63mer v2.04[31] by combining BAC and genome re-sequencing data. The N50 scaffold size was 0.37 kb. Given the low N50, we retained only those scaffolds over 1 kb (55 Mb) as a partial reference sequence. Genome re-sequencing data from 30 individuals were aligned to this reference using Bowtie2 v2.0.0-beta7 (ref. 32) with parameter–very-sensitive-local and then were re-ordered and sorted by Picard v1.84 (http://picard.sourceforge.net). Realigner TargetCreator and IndelRealigner[33] in GATK v2.1 were used to realign indels and UnifiedGenotyper[34] was used to call genotypes across 30 individuals using the following parameters: heterozygosity 0.01, stand_call_conf 50, stand_emit_conf 10, dcov 250. SNPs that were supported by more than 20 individuals, and with good quality ($Q > 30$), were used in the subsequent analysis, which yielded a total of 675,526 SNPs. We calculated F_{ST} between *cyrus* and *polytes* butterflies for each site using vcftools v0.1 (ref. 35). Scaffolds with the highest F_{ST} values were annotated and described based on BLASTn and BLASTx searches against NCBI nt and nr databases (Extended Data Table 2). The top candidate contained *dsx* female exon 4.

Reference sequence for local association mapping. SeqMan NGen (DNASTAR) was used to perform reference-guided assembly of the *P. polytes dsx* region, using all available sequence data (BAC and re-sequencing data) and an unpublished *Papilio xuthus* scaffold as a template. The resulting targeted assembly consisted of a single 4-Mb scaffold, which we used as a reference for local SNP calling, following the workflow outlined above for the GWAS analysis. We calculated false-discovery rate corrected *P* values[36] using PLINK v1.07 (ref. 37) to examine the strength of association between genotype and phenotype for each SNP across the 4-Mb scaffold.

HKA test. We used the HKA test[38] to compare the level of polymorphism in *dsx* to neighbouring genes. This test requires sequence data from another species to calculate interspecific divergence, for which we used an assembled transcriptome from *Papilio canadensis*[39]. We did not include the gene *sir2* in this analysis because the ORF recovered from *P. canadensis* was too small.

Linkage disequilibrium analysis and structural variant detection. Pairwise linkage disequilibrium was estimated using PLINK v1.07 (ref. 37). We examined linkage disequilibrium using all biallelic SNPs in the first 1 Mb and the last 2 Mb of our 4 Mb *dsx* scaffold; *dsx* is located in the first 1 Mb. Linkage disequilibrium analyses were performed separately for *polytes* and *cyrus* groups, as well as a combined analysis. Sample PR370 was removed from the *polytes* group because it appeared to be heterozygous at the mimicry locus based on *dsx* SNP genotypes.

Structural variation was examined using Pindel v 0.2.5a3 (ref. 40). We focused on the region surrounding *dsx* and large structural variants. We considered structural variants between 500–517,888 bp, supported by at least five individuals, with end positions not located in transposable elements (identified using BLASTnt). Three inferred inversions were located in the region near *dsx*, two small inversions contained entirely in introns (417102–445257, 465077–472881) and one large inversion spanning the length of *dsx* (417081–512851). We then BLASTed scaffolds from a *de novo* assembly of the 15 re-sequenced *polytes* samples against our reference-guided assembly to identify scaffolds that matched sequence on both ends of *dsx*, indicating that they may span an inversion breakpoint. Eight scaffolds had well-supported partial matches to both ends of *dsx*. Extended Data Table 4

gives the BLAST details of these scaffolds. Each scaffold hits in two places when pairwise BLASTed against the 4-Mb reference-guided assembly of the *dsx* region (an assembly in which *dsx* is in the standard, non-inverted orientation). The table shows where each scaffold hits the assembly (query start and query end) and which parts of the scaffold hit (subject start and subject end). Importantly, each scaffold hits before and after *dsx* and the hits are in opposite orientations (clarified by the start and end base-pair positions listed under subject start and subject end). This is what we predict if an inversion has brought sequence from downstream of *dsx*'s 3′ end around to connect to sequence upstream of the 5′ end. Because the regions before and after *dsx* contain partially repetitive sequences, we believe that these scaffolds may be pointing to the same inversion, although we cannot rule out more complex structural variation in the area. The results suggest a 5′ breakpoint between 392890–419929 and a 3′ breakpoint between 510753–539234.

We designed PCR primers for partially overlapping products spanning the potential 5′ and 3′ inversion breakpoints (5_1F CCTGCTACTCTGTGCGCAC, 5_1R CAGTATGTCTGAGAATTCGCTAC; 5_2F GTAGCGAATTCTCAGAC ATACTG, 5_2R GAAGCCTCGGGACTGTAAC; 5_3F GTTACAGTCCCGAG GCTTC, 5_3R CCATGTCGTAGTCATAGCGAG; 5_4F CTCGCTATGACTAC GACATGG, 5_4R GCTCGCGATTCCGTCCG; 5_5F CGGACGGAATCGCG AGC, 5_5R CCAGAATGACTGCATTGATCTG; 3_1F CGTAACGAATACGCC GAC, 3_1R GTATGAAAGTGAATAGGGTTAGG; 3_2F CCTAACCCTATT CACTTTCATA, 3_2R CCTCTTTGTAATAGGCAATCGTGG; 3_3F CCACGA TTGCCTATTACAAAGAGG, 3_3R GACAATAGACATTTGATCTTGTGTG; 3_4F CACACAAGATCAAATGTCTATTGTC, 3_4R GAGTACAGATTTAGT ACAGATTGTAATG; 3_5F CTCTCAGAACTGAGGTTCTGTAGC; 3_5R GAT GTGATGAACTGAGAGTTTCCAG; 3_6F CTGGAAACTCTCAGTTCATCAC ATC, 3_6R GAACGCGAGTTCTCCCTTTG; 3_7F CAAAGGGAGAACTCGC GTTC, 3_7R GCACCGACTATGTTCCGTGTAG) and tested them using 10 homozygous *cyrus* females and 10 homozygous *polytes* females. These 20 females were different from those that were sequenced, and because *dsx* is autosomal, 10 females per group represent 20 tested haploid genomes for each mimicry phenotype. PCR tests of the 5′ breakpoint were uninformative because they either yielded no products (5_1, 5_2, 5_3) or products in all 20 samples (5_4 and 5_5). However, two overlapping PCR primer pairs on the 3′ side produced PCR products in all 10 *cyrus* females and no *polytes* females (Extended Data Fig. 4), isolating the 3′ breakpoint to an interval spanning 524401– 526499. This matches very well with the genomic interval of perfect SNP associations (Fig. 2c) which spans 405767–526117. As a whole, the data suggest there is an inversion polymorphism spanning approximately 405000– 526000, which contains just *dsx*.

Transcriptome assembly. RNA-seq data were generated from wing-disc-derived RNA from two laboratory-reared *P. polytes* pupae, one male and one female. RNA extraction, library preparation and Illumina sequencing followed standard protocols[39]. *De novo* transcriptome assembly and best ORF prediction were performed using Trinity 2012-06-08 (ref. 41). We annotated our 4-Mb targeted reference using Blat[42] to search against assembled transcriptomes.

cDNA sequencing. We dissected developing wing discs from four *P. polytes* pupae six days after pupation (approximately half way through the pupal stage): one male and one female homozygous for the *polytes* allele and one male and one female homozygous for the *cyrus* allele. We extracted total RNA using Trizol and generated cDNA using the Bio-Rad iscript cDNA synthesis kit. We then PCR amplified from cDNA the 3′ portion of *dsx* that is alternatively spliced using two different primer pairs (F1 GTCCTGGTCATACTTAATTA, R1 CTATTAGGTACTAAGT AAATC; F2 CAGACTACTGGAGAAGTTCC, R2 CTATACAGATCTAACACT AAG). PCR products were cloned using an Invitrogen TOPO TA cloning kit, Sanger sequenced, and aligned to full-length *dsx* transcripts from the *de novo* transcriptome assemblies. We generated a total of 16–23 sequences per sample, comparison of which revealed consistent isoforms between females and the same male isoform in males.

qRT–PCR. We performed two separate qRT–PCR experiments. First, we examined expression of the three female *dsx* isoforms in males and females, comparing expression in wing tissue and body (abdomen) tissue. For this experiment, we used butterflies that were heterozygous at the mimicry locus. Developing forewing discs, hind wing discs, and abdomen tissue were dissected out of developmental stages spanning 4th instar larvae to 72 h after pupation. Total RNA was extracted from each tissue using Trizol, from which we generated cDNA using the Bio-Rad iscript cDNA synthesis kit. qRT–PCR primers were designed to target products $<$200 bp from each female isoform (Isoform1F CGTCGCGGAAGATAGATGA AG, Isoform1R ATTCGTACGGAGTCCACTAATTG; Isoform2F CCGTTAGT CCTGGTCATACTT, Isoform2R TTCTGTATTAAAGTCCATACTGGC; Isoform3F CAGAAAATGCTGAGCGAAAT, Isoform3R CGATAATGCACGGCACAGCAC) and standard curves were generated for all primer pairs to estimate efficiency. We also analysed expression of *ef1-α* as a reference gene to normalize expression. Reactions were run on a Bio-Rad real-time CFX96 detection system using ABI's sybr green master mix. We then quantified and compared expression levels using the $2^{-\Delta\Delta CT}$ method. For plotting and analysis, samples were grouped into 'early' (4th and 5th instar larvae), 'mid' (1–2 day pupae) and 'late' (3–5 day pupae) developmental time points. Only results for forewing and abdomen are shown because hindwing expression mirrors forewing expression. Expression was visualized relative to maximal expression, which was set = 1 for each isoform. Expression variation was compared using ANOVA, which resulted in significant effects of sex, tissue and developmental time point ($P < 0.05$) and significant interactions ($P \leq 0.01$) for each isoform. We used Tukey's HSD test to compare means across treatments. As a second experiment, we compared expression of female isoforms 1 and 2 between mimetic and non-mimetic female wings using the same methods outlined above. This experiment had slightly different sampling so the groupings for analysis consisted of 'early' (5th instar larvae and $<$1 day pupae), 'mid' (1-3 day pupae) and 'late' (5–12 day pupae).

Protein structure prediction. We used Phyre2 (ref. 43) to predict secondary and tertiary structures of the *polytes* and *cyrus* Dsx proteins (female isoforms 1), as well Dsx from *Bombyx mori* (GenBank BAB19781.1). The Phyre2 predictions suggest that *B. mori* and *cyrus* proteins have similar alpha helix percentages whereas *polytes* is lower (*Bmdsx*, 44%; *cyrus*, 46%; *polytes*, 40%), because several predicted alpha helix motifs have been broken into shorter segments by substitutions in *polytes*. Similarly, *B. mori* and *cyrus* proteins show similar compact tertiary structures whereas *polytes* has a loose terminus.

Dsx immunohistochemistry. Immunodetection of Dsx was carried out using a monoclonal anti-*Drosophila* Dsx DM domain (DNA-binding domain) antibody[44] (1:100; gift of C. Robinett), a secondary goat anti-mouse IgG antibody (1:250; Jackson ImmunoResearch) and a tertiary anti-goat-AlexaFluor555 antibody (1:250; Jackson ImmunoResearch).

31. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18 (2012).
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
33. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
34. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
35. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
36. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1188 (2001).
37. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
38. Hudson, R. R., Kreitman, M. & Aguade, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).
39. Zhang, W., Kunte, K. & Kronforst, M. R. Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using *de novo* transcriptome assemblies. *Genome Biol. Evol.* **5,** 1233–1245 (2013).
40. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
41. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol.* **29,** 644–652 (2011).
42. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
43. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4,** 363–371 (2009).
44. Mellert, D. J., Robinett, C. C. & Baker, B. S. *doublesex* functions early and late in gustatory sense organ development. *PLoS ONE* **7,** e51489 (2012).