Cell Reports

Supplemental Information

Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Defense

Qian Cong, Dominika Borek, Zbyszek Otwinowski, and Nick V. Grishin

Tiger Swallowtail genome reveals mechanisms for speciation and caterpillar defense

Qian Cong[2], Dominika Borek[2], Zbyszek Otwinowski[2], and Nick V. Grishin[1,2]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA. [2]Department of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816, USA.

Index for supplementary materials

Supplementary Figures and Figure Legends

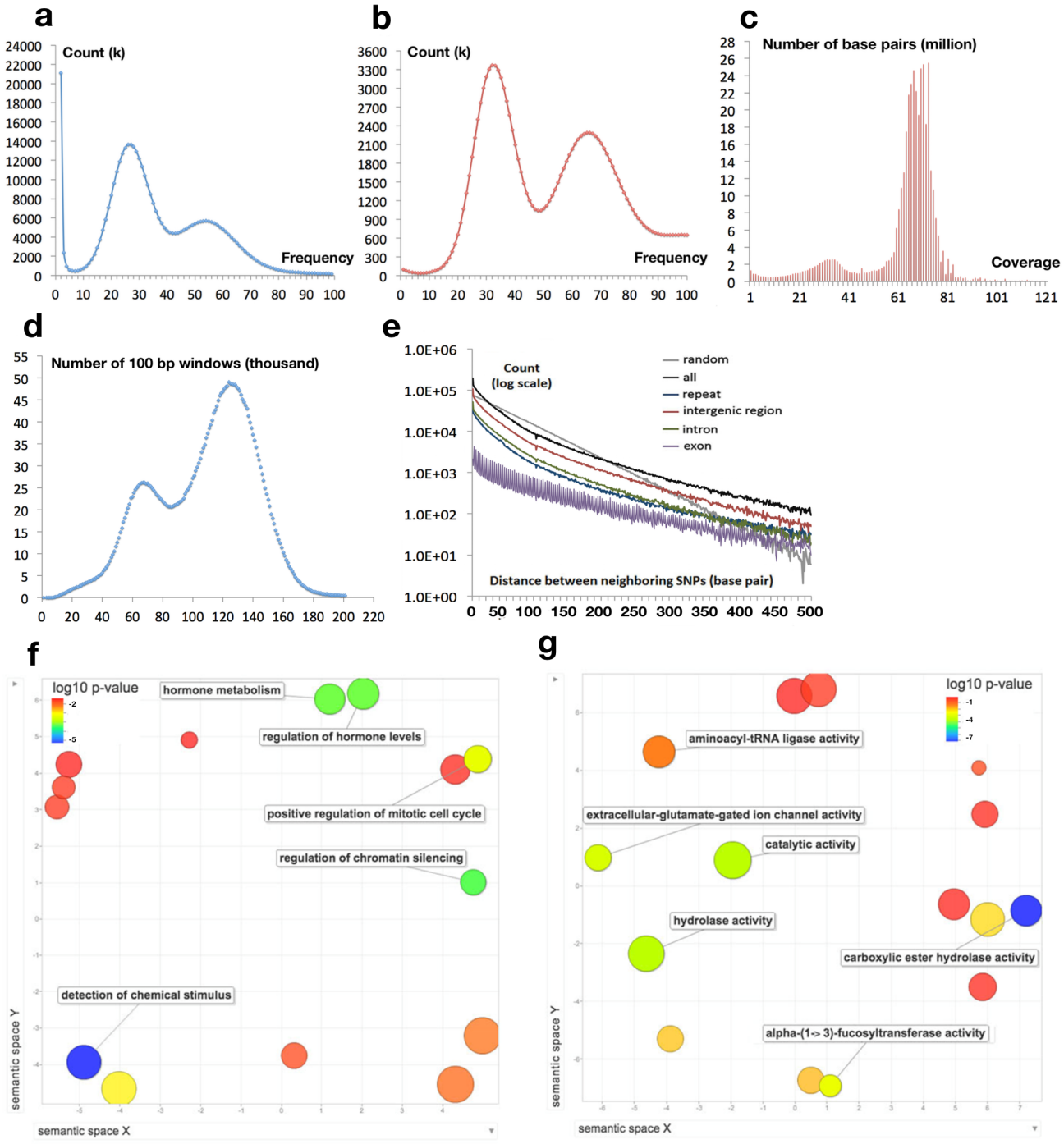

Figure S1. Essential plots for genome assembly, distribution of SNPs in the genome and function of SNP-enriched proteins, Related to Figure 1.

(a) Histogram of 19-mer frequencies in the reads from all the sequencing libraries before error correction by QUAKE. This graph is used to estimate the cutoff for k-mer based error correction, and it shows that 19-mers with frequency less than 7 are dominated by sequencing errors.

(b) Histogram of 15-mer frequencies in the reads for *Papilio glaucus* genome assembly (after error correction by QUAKE). This graph is used to estimate the genome size. The reads after data processing have an average length of 137 bp and a total length of 27,011,000,000 bp. This graph shows that 15-mers from homozygous regions are covered approximately 65 times. Therefore, the estimated coverage for the genome is: coverage estimated from k-mer frequency x average read length / (average read length – k-mer length + 1) = 65x137/(137-15+1) = 72.4 fold[6], and the estimated genome size is 27,011,000,000/72.4 = 373 million base pairs.

(c) Average coverage for scaffolds from the *Papilio glaucus* assembly_V0 (the result directly obtained from the Platanus assembler). The peak at 34 fold coverage is likely dominated by short scaffolds originating from highly heterozygous regions that were not merged to the equivalent segments in the homologous chromosomes.

(d) The numbers of reads mapped to 100 bp non-overlapping windows from the *Papilio glaucus* genome assembly_V0 (direct result from the Platanus assembler). The peak on the left represents highly heterozygous regions in the genome where the equivalent regions in the homologous chromosomes cannot be merged by the assembler due to low levels of sequence similarity.

(e) Non-random distribution of SNPs in different genomic regions revealed by the distribution of distances between neighboring SNP pairs. Since the Y-axis (count) is log scale, a random distribution of SNPs will show a constant slope as the random control in grey. The SNPs in the whole genome, intergenic regions, introns, exons and repeats are all distributed non-randomly: they tend to concentrate in certain regions and avoid others. The purple line for SNPs in exons displays large fluctuations because SNPs in the protein coding regions happen at the third position of codons more than other positions; therefore, the neighboring SNPs are more likely to be separated by 3N-1 base pairs.

(f)-(g), Significantly enriched GO terms associated with *Papilio glaucus* proteins that are significantly enriched in SNPs in the categories of "biological process" and "molecular function", respectively. All these GO terms are listed in Supplementary Table S8. Each dot represents a family of GO terms (children are merged into parental GO terms) and its size correlates with the number of GO terms in this family. The color of the dots shows the significance level of their enrichment as labeled in the upper left corner. The distances between dots represent the similarity in their meanings as defined by REVIGO web server. We showed the annotation for the most significantly enriched GO terms with Q-values below 0.2 in False Discovery Rate tests.

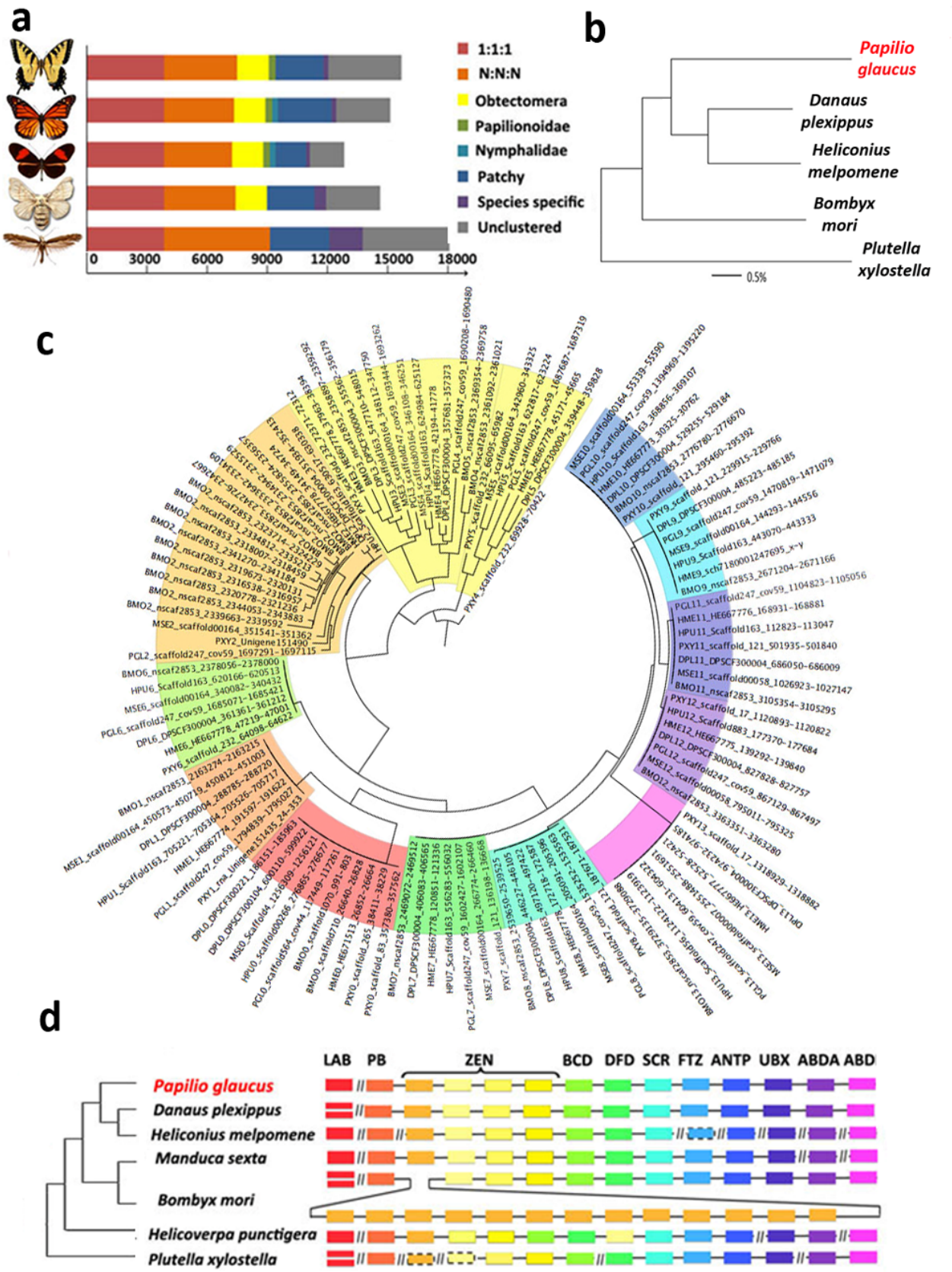
a
1:1:1
N:N:N
Obtectomera
Papilionoidae
Nymphalidae
Patchy
Species specific
Unclustered
0
3000
6000
9000
12000
15000
18000
b
Papilio glaucus
Danaus plexippus
Heliconius melpomene
Bombyx mori
Plutella xylostella
0.5%
c
d
LAB
PB
ZEN
BCD
DFD
SCR
FTZ
ANTP
UBX
ABDA
ABD
Papilio glaucus
Danaus plexippus
Heliconius melpomene
Manduca sexta
Bombyx mori
Helicoverpa punctigera
Plutella xylostella

Figure S2. Phylogeny of Lepidoptera species and synteny of Hox genes, related to Figure 2.

(a) Number of different types of orthologous groups in each Lepidoptera species with published genomes (on the right) and evolutionary tree (on the left) for them based on the concatenated alignment of universal single-copy orthologs (1:1:1) that. N:N:N: multiple-copy (more than one in at least one species) orthologs shared among all species; Obtectomera: orthologs specific to Obtectomera (all other four species except *Plutella xylostella*); Papilionoidae: orthologs specific to Papilionoidae (*Papilio glaucus*, *Danaus plexippus* and *Heliconius melpomene*); Nymphalidae: orthologs specific to Nymphalidae (*Danaus plexippus* and *Heliconius melpomene*); Patchy: orthologs that are shared between more than one, but not all species; Specific: specific to only one species and that have close homologs within that species; Unclustered: proteins that are not in any orthologous groups.

(b) Phylogenetic tree of the *Lepidoptera* species by the rate of gene inversions, an indicator of synteny. *Papilio glaucus* is placed together with other butterflies in this tree.

(c) Phylogenetic tree of the Hox genes from Lepidoptera species. Seven species were included in this tree, i.e. *Bombyx mori* (BMO), *Plutella xylostella* (PXY), *Heliconius melpomene* (HME), *Danaus plexippus* (DPL), *Papilio glaucus* (PGL), *Manduca sexta* (MSE) and *Helicoverpa punctigera* (HPU). Each node in the tree is labeled by the abbreviation of the species name and the No. of orthologous groups to which this protein belongs, followed by the scaffold and range on the scaffold to which the homeodomain map. Equivalent homeodomains are highlighted in the same color: red: LAB, orange: PB, yellow orange and yellow: ZEN-like, lime: BCD, green: DFD, cyan: SCR, light blue: FTZ, blue: ANTP, dark blue: UBX, purple: ABDA, pink: ABDB.

(d) Arrangement of Hox genes on genome scaffolds from all Lepidoptera species with available genome sequences. Equivalent genes are shown as boxes of the same color and the corresponding *Drosophila* Hox gene names are labeled above. Double boxes in the same position indicate gene duplications and "//" marks the boundaries between different scaffolds. The genome sequences of *Manduca sexta* and *Helicoverpa punctigera* are deposited in the database but are not yet published. *Helicoverpa punctigera* has a unique gene inversion in these Hox genes, and the additional expansion of *Zen*-like Hox genes is unique to *Bombyx mori.*

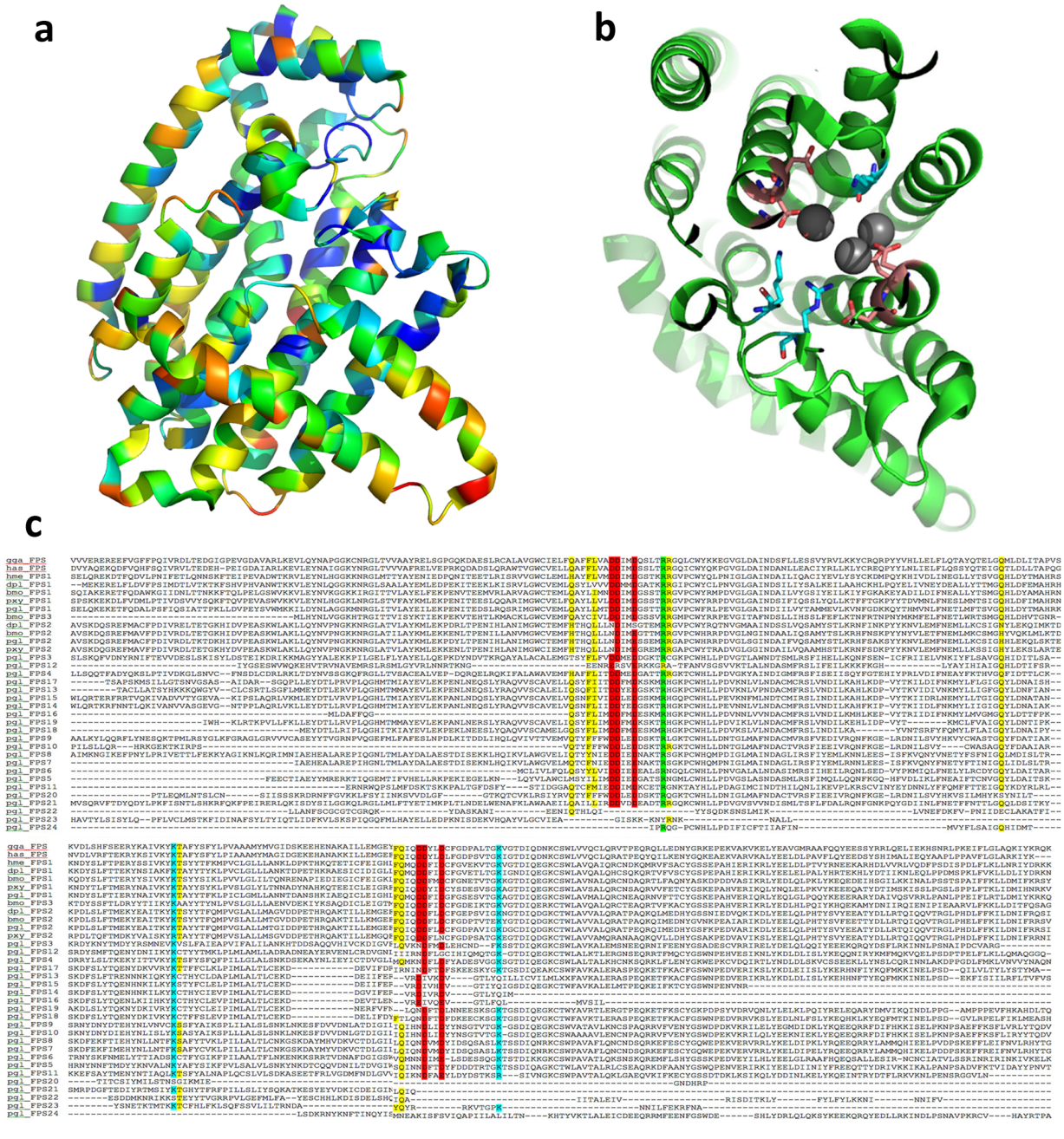

Figure S3. Structure and alignment of Farnesyl Pyrophosphate Synthase (FPS) homologs, related to Figure 2.
(a) Structure template (human FPS) for FPS homologs adopts a typical isprenoid synthase fold. The structure is olored in rainbow according to the conservation level in the alignment shown in Fig. S3c (the most conserved residues in blue and the least conserved residues in red). PDB ID: 4GA3. **(b) Structure template for FPS homologs with the catalytic residues highlighted.** The three magnesium ions in the catalytic site are colored in grey; negatively charged residues used to coordinate these magnesium ions are highlighted in pink and the positively charged residues pointed to the catalytic sites are shown in cyan. PDB ID: 4GA3. **(c) Alignment of the FPS homologs with functional sites highlighted.** The positions colored in red and cyan are conserved charged residues that contribute to the catalytic sites, and the one colored green has been shown to be very important in mutagenesis experiments. The residues lining the substrate-binding pocket are highlighted in yellow. All these functionally relevant positions, residues with different properties from the human sequence are not highlighted.

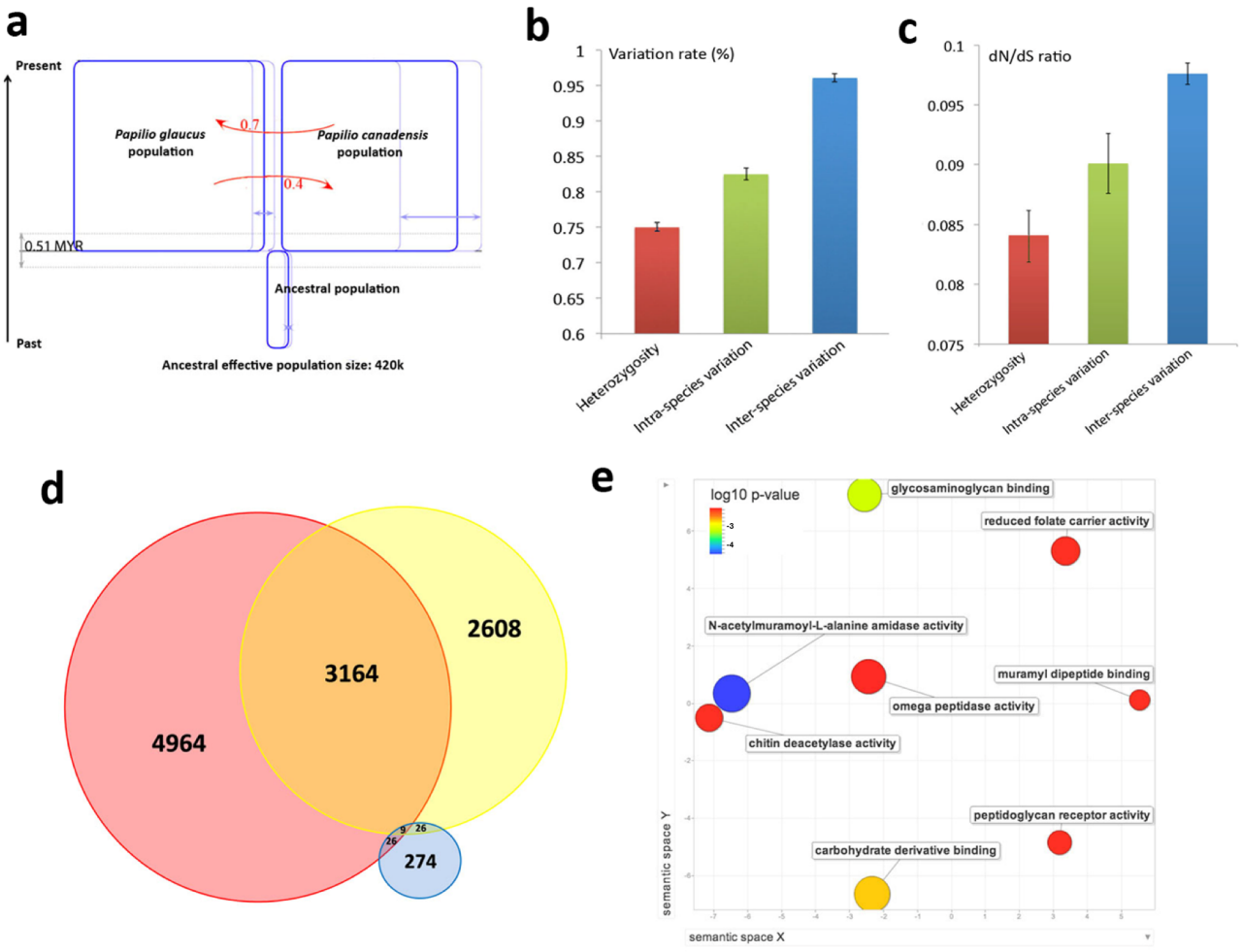

Figure S4. Speciation between *Papilio glaucus* (*Pgl*) and *Papilio canadensis* (*Pca*), related to Figure 3.
(a) Isolation with migration model[88,90] for speciation between *Pgl* and *Pca*. The widths of the three blue boxes are correlated with the estimated effective population sizes. The light blue lines show the confidence intervals (95%). The vertical dimension corresponds to time in evolution. The time (0.51 million years ago) at which the two species to diverge is marked by the grey solid line and the grey dashed lines indicate the 95% confidence interval. The values by the red arrows indicate the estimated migration rates (migration per mutation). **(b) The overall mutation rate between *Pgl* and *Pca* (blue column) is significantly higher (confidence level > 99%) than the overall mutation rate within species (green column); (c) The overall dN/dS ratio between *Pgl* and *Pca* (blue column) is significantly higher (confidence level > 95%) than the dN/dS ratio within species (green column).** In both **(b)** and **(c),** the red column shows statistics for heterozygosity in the *Pgl* specimen with its whole genome sequenced. The error bars for intra- and inter- species variations indicate the standard deviations of the same measurements on different pairs of specimens, and the errors for heterozygosity is estimated from the standard deviation of heterozygosity on individual proteins. **(d) Venn diagram of different classes of positively selected positions in *Pgl* and *Pca* detected by PAML.** Red: positions with non-synonymous mutations within the *Pgl* specimens; Yellow: positions which show non-synonymous mutations within the *Pca* specimens; Blue: positions which show non-synonymous mutations between *Pgl* and *Pca*. Among the positions that underwent positive selection, there are few (274, 2.5%) positions that are conserved within both species but differ between *Pgl* and *Pca*. **(e) The significantly enriched GO terms in the category of "molecular function" associated with divergence hotspots.** All of these GO terms are listed in Supplemental Table S24. Each dot represents a family of GO terms (children are merged into parental GO terms) and its size correlates with the number of GO terms in this family. The color of the dots shows the significance level of their enrichment as shown in the label at the upper left corner. The distances between dots represent the similarity in their meanings as defined by REVIGO web server. We showed the annotations for the most significantly enriched GO terms with Q-values below 0.2 in False Discovery Rate tests.

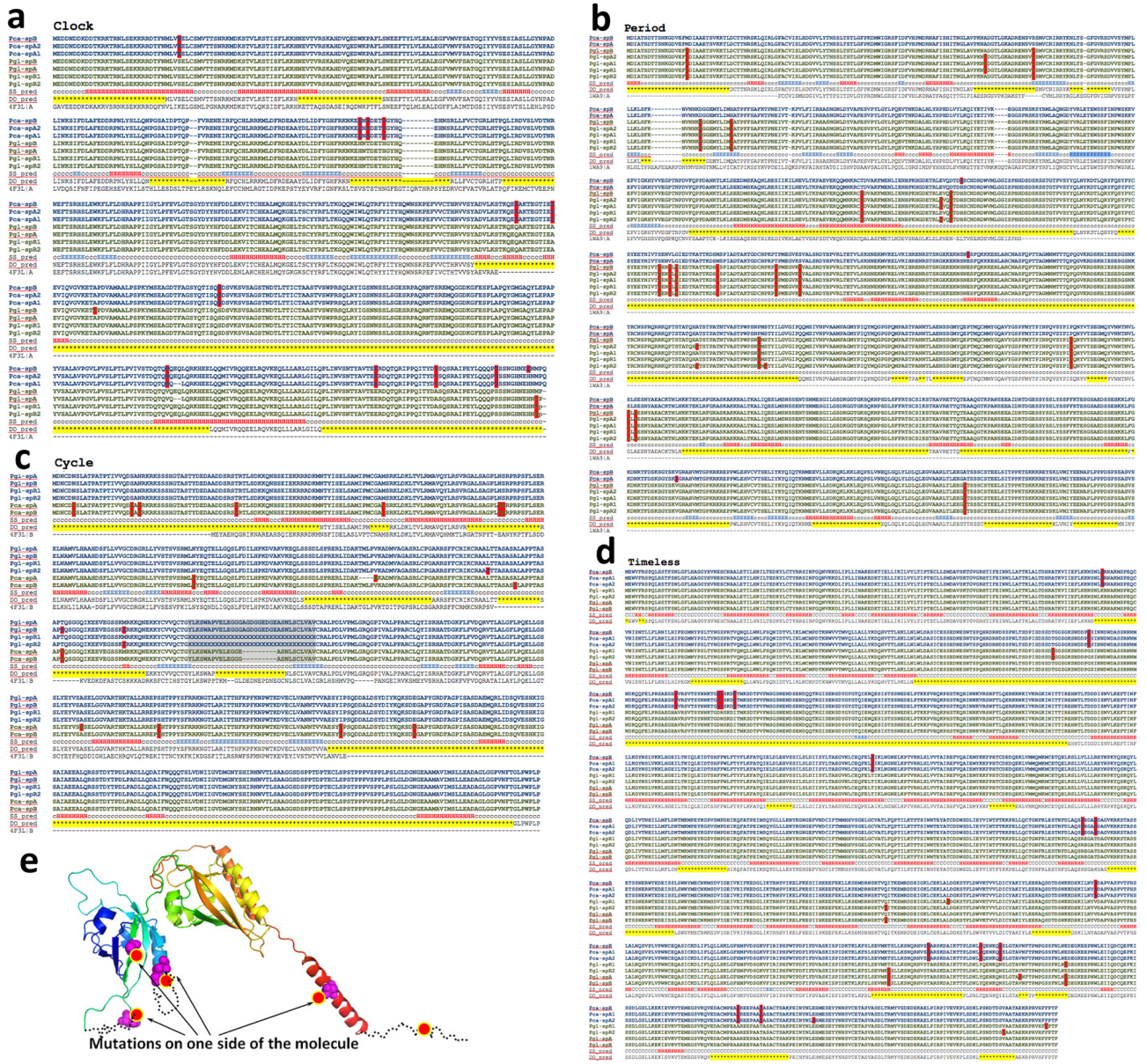

Figure S5. Sequence analysis of Circadian Clock Proteins, related to Figure 4.
(a)-(d) Multiple sequence alignment, secondary structure and disordered region prediction, and alignment to a 3D structure template for the protein CLOCK, PEROID, CYCLE and TIMELESS, respectively. There is no confident structure template for TIMELESS. Positions with intra- or inter-species variations are highlighted in red (the amino acid with lower frequency is highlighted); SS_pred: secondary structure prediction, and in this line, H indicates helix, E means strand, C means coil or loop. DO_pred: disordered region prediction, and the predicted disordered positions are marked by * and highlighted in yellow. **(e) Three dimensional structure template of the protein PERIOD with positions that are different between *Papilio glaucus* and *Papilio canadensis* mapped.** The structure template is the protein PERIOD from *Drosophila* (PDB code: 1WA9). Disordered regions are represented by black dots. The side chains of residues that differ between *Papilio glaucus* and *Papilio canadensis* are shown in magenta and marked by red dots. All inter-species mutations are on the same side of the molecule.

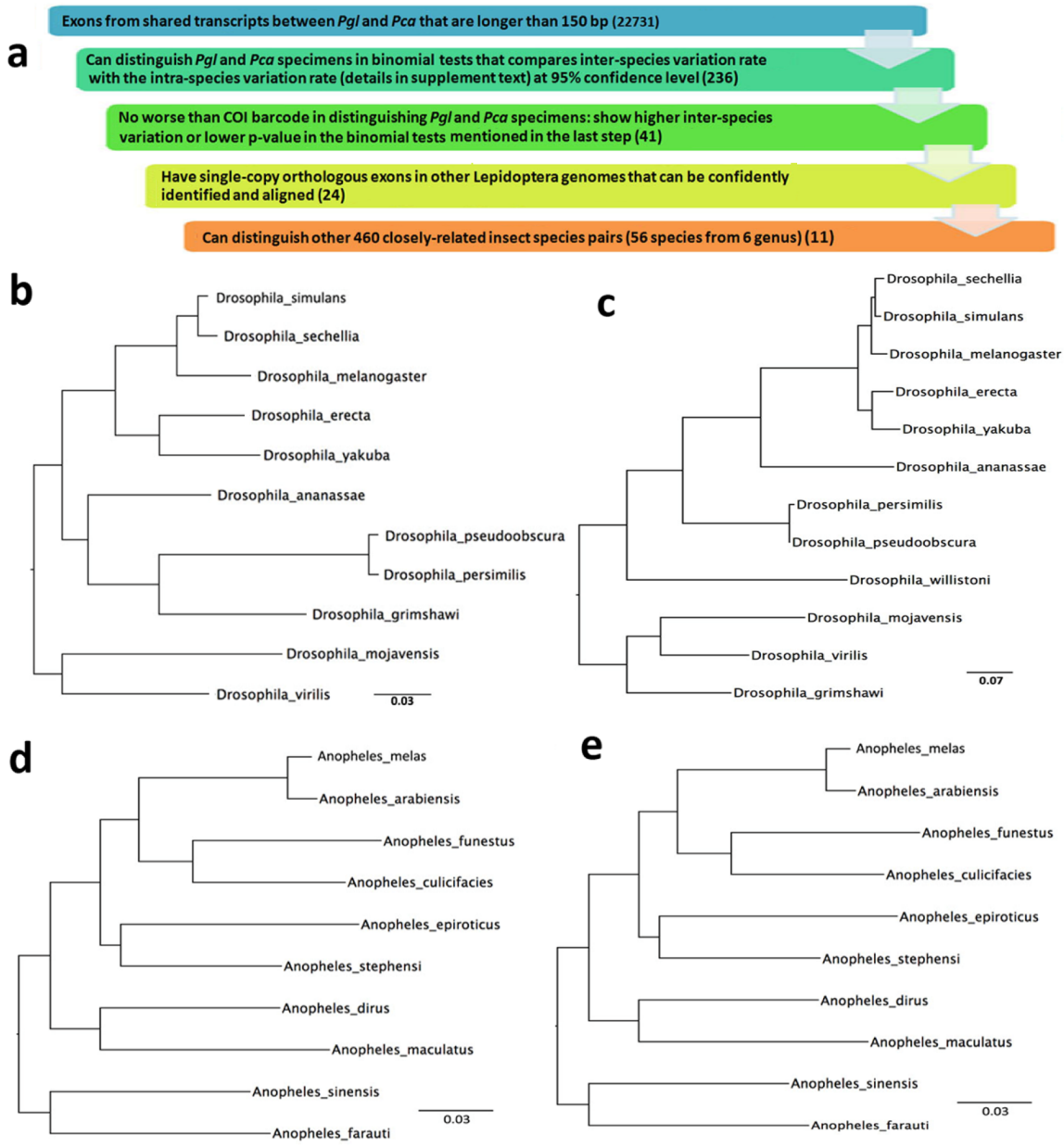

Figure S6. Workflow for nuclear DNA barcode selection and the ability for the barcodes to reflect the phylogeny of *Drosophila* and *Anopheles* species, related to Table 2.

(a) Workflow for selecting nuclear DNA barcodes to distinguish closely related insects. The number of remaining barcode candidates after each step is show in parentheses. **(b) Phylogenetic tree of *Drosophila* species based on COI barcode sequences.** The tree topology mostly agrees with the phylogenetic tree based on whole genome data in FlyBase (http://flybase.org/static_pages/species/sequenced_species.html) except for the placement of *Drosophila grimshawi.* **(c) Phylogenetic tree of *Drosophila* species based on proposed nuclear barcode pgl6110.7_exon1.** The tree topology entirely agrees with the phylogenetic tree based on whole genome data in FlyBase. **(d) Phylogenetic tree of *Anopheles* species based on COI barcode sequences.** The tree topology is very different from the one based on whole genome data in VectorBase. (https://agcc.vectorbase.org/index.php/Main_Page). **(e) Phylogenetic tree of *Anopheles* species based on proposed nuclear barcode pgl6877.3_e2.** The tree topology completely agrees with the one based on whole genome data in VectorBase.

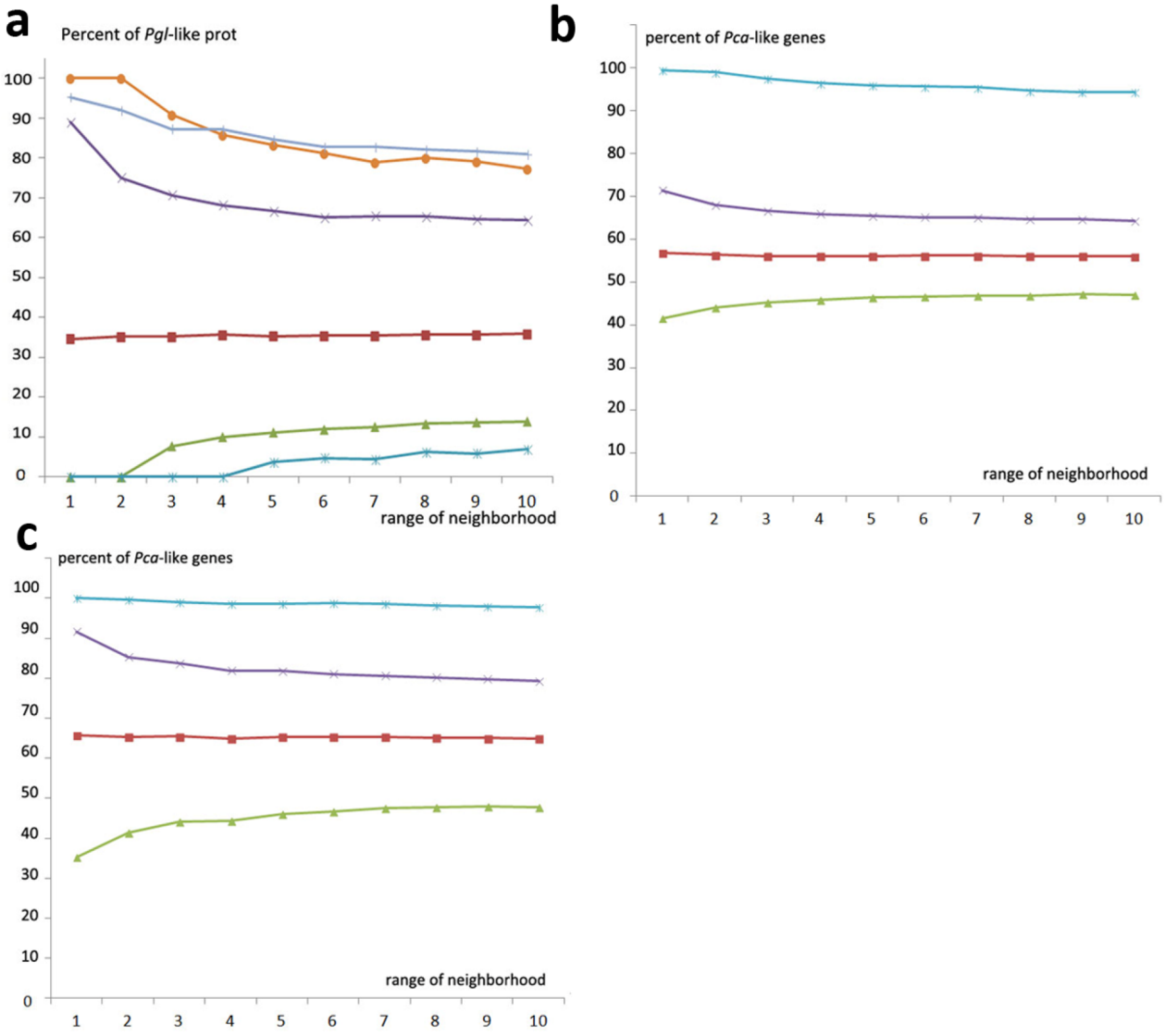

Figure S7. Clustering of *Papilio glaucus* (*Pgl*)-like and *Papilio canadensis* (*Pca*)-like proteins from *Papilio appalachiensis* (*Pap*) in the genome, related to Figure 5.

(a) Percentage of *Pgl*-like proteins (criterion: difference between "average identity to *Pgl* sequences" and "average identity to *Pca* sequences" is larger than 0.6%) in the neighborhood of statistically supported *Pgl*-originated proteins is significantly (confidence level >98.1%) higher than those in the neighborhood of randomly selected samples. Blue: confident *Pgl*-originated transcripts; orange: best 0.1% of random sets; purple: best 1.9% of random sets; red: average for random sets; green: worst 1.9% of random sets; cyan: worst 0.1% of random sets. **(b) Percentage of *Pca*-like proteins (criterion: difference between "average identity to *Pca* sequences" and "average identity to *Pgl* sequences" is larger than 0.2%) in the neighborhood of statistically supported *Pca*-originated proteins is significantly (confidence level >99.99%) higher than those in the neighborhood randomly selected samples.** Cyan: confident *Pca*-originated transcripts; purple: best 0.01% of random sets; red: average for random sets; green: worst 0.01% of random sets. **(c) Percentage of *Pca*-like proteins (criterion: difference between "average identity to *Pca* sequences" and "average identity to *Pgl* sequences" is larger than 0.6%) in the neighborhood of statistically supported *Pca*-originated proteins is significantly (confidence level >99.99%) higher than those in the neighborhood of randomly selected samples.** Cyan: confident *Pca*-originated transcripts; purple: best 0.01% of random sets; red: average for random sets; green: worst 0.01% of random sets.

Supplementary Table Legends

Table S1. Statistics for sequencing data, data processing, genome assembly and genome annotation, related to experimental procedure.

Table S1A. Amount of data in sequence libraries used for genome assembly

Table S1B. Amount of data that is removed at each stage of data processing

Table S1C. Assembly statistics for *Papilio glaucus* genome by different methods
* assembly_v1: the current assembly, improved based on Platanus assembly

Table S1D. Repeats in the *Papilio glaucus* genome

Table S1E. Methods used for protein coding gene annotation

Table S1F. All predicted proteins in the *Papilio glaucus* genome and their function annotations

Table S2. Quality and composition of Lepidoptera genomes, related to Table 1.

Table S2A. Information about other published Lepidoptera genomes

Table S2B. Composition of Lepidoptera genomes

Table S2C. Comparison of CEGMA gene coverage by a single scaffold in Lepidoptera genome drafts

Table S3. Distribution of SNPs in the *Papilio glaucus* genome, related to Figure 1.

Table S3A. SNP rate in different regions of the *Papilio glaucus* genome

Table S3B. Protein coding genes that are significantly enriched in SNPs

Table S3C. Enriched GO terms associated with protein coding genes that are significantly enriched in SNPs

Table S4. Comparative analyses of the Lepidoptera genomes, related to Figures 2.

Table S4A. Synteny between all published Lepidoptera species measured by percent of genes in micro-syntenic blocks.
strict microsynteny: requiring one gene to be in microsyntenic block on both upstream and downstream sides
microsynteny: requiring one gene to be in syntenic block with at least one neighboring gene

Table S4B. Sequences for all the homeodomains detected in Lepidoptera genomes

Table S4C. Expanded protein families in *Papilio glaucus*
Gene expansion index 1: ratio of number of *Papilio glaucus* proteins and average number of proteins from other species in the orthologous group
Gene expansion index 2: ratio of total length of *Papilio glaucus* proteins and average total length of proteins from other species in the orthologous group

Table S4D. Opsins from all published Lepidoptera genomes

Table S4E. Eclosion hormone molecules from all published Lepidoptera genomes

Table S4F. Farnesyl pyrophosphate synthase homologs from all published Lepidoptera genomes

Table S4G. Number of farnesyl pyrophosphate synthase homologs detected in transcriptome of other swallowtail species

Table S5. Speciation between *Papilio glaucus* and *Papilio canadensis*, related to Figure 3.

Table S5A. Results of isolation with migration models built on different sets of shared transcripts between *Papilio glaucus* and *Papilio canadensis* specimens

Table S5B. Intra- and inter- species variation rates based on the 8230 transcripts from the 3 *Papilio glaucus* and 2 *Papilio canadensis* specimens

Table S5C. Speciation hotspots: genes that always show larger inter-species variation on both the protein and DNA sequence level

Table S5D. Proteins with positively selected sites in the *Papilio glaucus* and *Papilio canadensis*
class I: same within each species, and different between species
class II: different between species
class III: different within *Papilio glaucus* specimens
class IV: different within *Papilio canadensis* specimens

Table S5E. Enriched GO terms associated with speciation hotspots between *Papilio glaucus* and *Papilio canadensis*
P value1: binomial test for the enrichment of the GO-term
P value2: binomial test for enrichment of inter-species mutations in proteins associated with these GO-terms

Table S6. Selection of nuclear DNA barcodes, related to Table 2.

Table S6A. Performance of commonly used nuclear DNA barcodes and new candidates on Lepidoptera genomes

Table S6B. Other insect genomes used to select and validate nuclear DNA barcodes

Table S6C. Additional information of selected nuclear DNA barcodes

Table S7. Genome confirms *Papilio appalachiensis* as a hybrid species, related to Figure 5.

Table S7A. Intra- and inter- species variation rates based on the 7410 transcripts from the 3 *Papilio glaucus*, 2 *Papilio canadensis* and 2 *Papilio appalachiensis* specimens

Table S7B. *Papilio appalachiensis* protein coding genes that originated from *Papilio glaucus*

Table S7C. *Papilio appalachiensis* protein coding genes that originated from *Papilio canadensis*

Table S7D. *Papilio appalachiensis* protein coding genes that are different from both *Papilio glaucus* and *Papilio canadensis*

Table S7E. Genes on the same scaffold as the 6PGD, a gene closely linked to melanic female regulating gene

Extended Experimental Procedures

S1 Sequencing library preparation protocol

S1.1 Genomic DNA extraction

A piece of muscle (about 80 mg) was extracted from the thorax of a freshly caught and frozen male *Papilio glaucus*, and the remaining almost intact specimen is preserved. Genomic DNA was extracted from this piece of tissue with ChargeSwitch gDNA mini tissue kit following the manufacturer's protocol with modifications.

A. Lysis

Divide the muscle into 4 pieces, and do the following steps for each piece (about 20 mg):

Cut the muscle thoroughly into less than $1mm^3$ pieces with a scalpel on a Petri dish;
Add 0.5 ml lysis buffer (L15) to the Petri dish and transfer the tissue to a 1.5 ml tube;
Wash the Petri dish with 0.5 ml lysis buffer and transfer the wash to the same tube;
Add 30 µl Proteinase K (20 mg/ml), vortex and incubate at 55 °C overnight;
Add 20 µl RNase A (5 mg/ml), vortex and incubate at room temperature for 10 minutes.

B. Bind DNA

Resuspend the magnetic beads and do the following steps:

Add 120 µl of Purification Buffer (N5) to each tube and vortex to mix;
Add 100 µl of Magnetic Beads and flip the tube up and down to mix;
Incubate at room temperature for 10 minutes;
Place the tube in the MagnaRack for 2 minutes;
Remove supernatant and discard.

C. Wash beads

Wash the beads twice and for each time:

Remove the tube from the MagnaRack;
Add 1 ml of Wash Buffer (W12);
Gently pipet up and down to resuspend the beads;
Place the tube in the MagnaRack for 2 minutes;
Remove the supernatant and discard.

D. Elute DNA

Elute the DNA from the beads three times to increase the yield, and for each time:

Remove the tube from the MagnaRack;
Add 100 µl of Elution Buffer (E5);

Pipet gently to resuspend the beads;
Incubate at 37°C for 5 minutes;
Place the tube on the MagnaRack for 2 minutes;
Transfer the supernatant containing the purified DNA to a clean tube.

E. Quantify the amount of DNA

Use the Qubit dsDNA HS Assay Kit and Qubit fluorometer to measure the concentration of DNA following the manufacturer's protocol. From 80 mg muscle, we obtained approximately 20 µg of DNA. Check the quality of genomic DNA using the E-gel 0.8% Agarose gel. One should expect to obtain long DNA fragments (about 40 kb) at this step, and this is necessary for the subsequent steps.

S1.2 Paired-end library preparation protocol

We prepared 250bp and 500bp paired-end libraries following a protocol similar to the Illumina TruSeq DNA sample preparation guide. For each paired-end library, approximately 1 µg genomic DNA was used.

A. Fragmentation

Material: Covaris S220 Focused-ultrasonicator, Covaris microTUBE and genomic DNA.
Parameters for the 250 bp library:

Intensity: 5
Duty cycle: 10%
Cycles per burst: 200
Treatment time: 90s
Volume: 50 µl
Temperature: 7°C

Parameters for 500 bp library:

Intensity: 5
Duty cycle: 5%
Cycles per burst: 200
Treatment time: 35s
Volume: 50 µl
Temperature: 7°C

B. End repair

Material: NEBNext End Repair Module and fragmented DNA.
Prepare the reaction in a 0.5 ml PCR tube:

50 µl fragmented DNA
35 µl sterile H_2O
10 µl End Repair Reaction Buffer (10X)

5µl End Repair Enzyme Mix.

Incubate at 20°C for 30 min and keep at 4°C for 30 min. Purify DNA with Ampure XP beads (1.8x volume) and elute 2 times in a total volume of 40 µl.

C. dA-tailing

Material: NEBNext dA-Tailing Module and end-repaired DNA.

Prepare the reaction in a 0.5 ml PCR tube:

40 µl end-repaired DNA
2 µl sterile H_2O
5 µl dA-Tailing Reaction Buffer (10X)
3 µl Klenow fragment

Incubate at 37°C for 30 min. Purify DNA with Ampure XP beads (1.8x volume) and elute 2 times in a total volume of 33 µl.

D. Adapters Ligation

Material: NEBNext Quick Ligation Module, Illumina TruSeq adapters and dA-tailed DNA. Adapters with different indices are needed for different libraries if they will be sequenced on the same lane.

Prepare the reaction in a 0.5 ml PCR tube:

33 µl dA-tailed DNA
2 µl TruSeq adapter
10 µl Quick Ligation Reaction Buffer (5X)
5 µl T4 Quick Ligase

Incubate at 20°C for 30 min. Purify DNA with Zymo DNA cleanup and concentrator-5 kit and elute twice in a total volume of 35 µl.

E. PCR amplification

Material: PCR Primer Cocktail, PCR Mater Mix from TruSeq DNA sample Prep V2 kit, and adapter-ligated DNA.

Prepare the reaction in a 0.5 ml PCR tube:

5 µl PCR primer cocktail
15 µl PCR master mix
35 µl adapter-ligated DNA

Do PCR in a thermal cycler with a heated lid using the following program:

98°C for 30s
8 cycles of:
    98°C for 10s
    60°C for 30s
    72°C for 30s
72°C for 5 min
Hold at 4°C

Purify DNA with Ampure XP beads (1.8x volume) and elute twice in a total volume of 40 µl.

F. Size selection

Material: E-Gel EX 2% Agarose gel with the E-gel base and Trackit 50 bp DNA ladder from Invitrogen and PCR amplified DNA.
Distribute the PCR product into 4 lanes and dissect the band (about 4mm) at the desired fragment size. Recover the DNA from the gel using the Zymoclean gel DNA recovery kit following the manufacturer's protocol and elute to a final volume of 20 µl to obtain the final library.

S1.3 Mate pair library preparation protocol

We prepared 2 kb, 6 kb and 15 kb mate pair libraries using a modified version of a previously published mate pair library preparation protocol[1]. For the 2 kb, 6 kb and 15 kb libraries, about 1.7 µg, 3 µg and 7.2 µg genomic DNA was used, respectively.

A. Fragmentation 1

Material: Covaris S220 Focused-ultrasonicator, genomic DNA, Covaris miniTUBE (white) and Covaris gTUBE.
For a 2 kb library, prepare 1-2 µg DNA in 200 µl solution and shear DNA with Covaris S2 equipment in Covaris miniTUBE (white) using the following parameters:

Temperature: 7°C
Duty factor: 20%
Peak incident Power: 3
Cycles per burst: 1000
Treatment time: 15 min

For a 6kb library, prepare 2-4 µg DNA in 150 µl solution and shear DNA using Covaris gTUBE at eppendorf 5415R centrifuge under the following condition:

Speed: 12000 rpm
Temperature: 20 °C
Treatment time: 30s and then flip the tube, treat for another 30s

For a 15 kb library, prepare 6-8 µg DNA in 150 µl solution and shear DNA using Covaris gTUBE and eppendorf 5415R centrifuge under the following condition:

Speed: 5500 rpm
Temperature: 20 °C
Treatment time: 1 min and then flip the tube, treat for another 1 min

Purify DNA with Ampure XP beads. For 2 kb fragment, add 140 µl beads (0.7x volume); for 6 kb and 15 kb fragments, add 75 µl beads (0.5x volume). Elute with Zymo Zyppy elution buffer twice at 37 °C for 10 min in a total volume of 85 µl (add 44 µl each time).

B. End Repair 1

Material: NEBNext End Repair Module and fragmented DNA.
Prepare the reaction in a 0.5 ml PCR tube:

85 µl fragmented DNA
10 µl End Repair Reaction Buffer (10X)

5µl End Repair Enzyme Mix

Incubate at 20°C for 30 min and keep at 4°C for 30 min. Purify DNA with Ampure XP beads (0.7x volume for 2 kb library and 0.5x volume for 6 kb and 15 kb libraries). Elute with Zymo Zyppy elution buffer twice at 37 °C for 10 min in a total volume of 42 µl (add 22 µl each time).

C. dA-tailing 1

Material: NEBNext dA-Tailing Module and end-repaired DNA.

Prepare the reaction in a 0.5 ml PCR tube:

42 µl end-repaired DNA
5 µl dA-Tailing Reaction Buffer (10X)
3 µl Klenow fragment

Incubate at 20°C for 30 min and keep at 4°C for 30 min. Purify DNA with Ampure XP beads (0.7x volume for 2 kb library and 0.5x volume for 6 kb and 15 kb libraries). Elute with Zymo Zyppy elution buffer twice at 37 °C for 10 min in a total volume of 60 µl (add 31 µl each time).

D. Ligation to circularization adapters

Material: NEBNext Quick Ligation Module and the circularization adapters with loxP sites (customized DNA oligos from Integrated DNA Technology):

loxP1 double-stranded DNA oligo:
forward strand (with biotin label):
5’ CGATAACTTCGTATAATGTATGCTATACGAAGT(Bio-dT)ATTACGT 3’
reverse strand (with 5’ phosphate):
5’ (5Phos)CGTAATAACTTCGTATAGCATACATTATACGAAGTTATCGACC 3’

loxP2 double-stranded DNA oligo:
forward strand (with biotin label):
5’ GCATAACTTCGTATAGCATACATTATACGAAGT(Bio-dT)ATACGAT 3’
reverse strand (with 5’ phosphate):
5’ (5Phos)TCGTATAACTTCGTATAATGTATGCTATACGAAGTTATGCACC 3’

Prepare the reaction in a 0.5 ml PCR tube (mix annealed loxP1 with annealed loxP2 first):

60 µl dA-tailed DNA
5 µl annealed loxP1 adapter (500 µM)
5 µl annealed loxP2 adapter (500 µM)
20 µl Quick Ligation Reaction Buffer (5X)
10 µl T4 Quick Ligase

Incubate at 20°C for 30 min. Purify immediately after the incubation (prevent further ligation that may lead to hybrid) with Zymo DNA cleanup and concentrator-5 kit following the manufacturer’s protocol and elute twice in a total volume of 36 µl (add 19 µl each time).

E. Size selection

Material: E-gel EX 1% Agarose gel with the E-gel base, TrackIt 1kb DNA ladder and adapter-ligated DNA.

Distribute the PCR product in 4 lanes and run the gel until the DNA ladder is well separated. Dissect the band (about 1 cm to include most of the long fragments) at the desired length. Recover the DNA from the gel using the Zymoclean gel DNA recovery kit (ADB buffer volume to

dissolve the gel: 1.5x for over 10 kb fragments and 2x for others) and elute twice to a final volume of 40 µl.

F. Concentration measurement

Material: Qubit dsDNA HS Assay Kit and Qubit fluorometer.
Follow the manufacturer's protocol to measure the concentration.
Note that in order to be successful with the following procedure, we recommend at least 400 ng DNA for the 2 kb library, 600 ng for the 6 kb library and 800 ng for the 15 kb library at this step. One can lose quite a lot of DNA during the processes above, so it is necessary to start with more DNA or to do the DNA purification steps with great care. Since we had a limited amount of DNA from only a piece of muscle of a single specimen, we performed the purification for each step very carefully (usually we bind 2-3 times and elute 3 times to increase the efficiency). We had about 40% yield for 2 kb and 6 kb libraries and 25% yield for the 15 kb library. In general, one can expect about 30% yield for libraries below 10 kb and less (15% - 20%) for longer ones.

G. Circularization and Digestion

Material: Cre recombinase (with buffer), Plasmid-Safe ATP dependent DNase (with 25 mM ATP) and *E coli*. Exonuclease I.
Dilute the adapter-ligated and size-selected DNA to a 2.5 ng/µl concentration. Assume the volume of DNA is 80x µL (x indicates an unknown number). Prepare the reaction in a 0.5 ml PCR tube (the final concentration of DNA in the reaction mix will be 2 ng/µl):

80x µl DNA
10x µl Cre recombinase buffer (10X)
10x µl Cre recombinase (1U/µl)

Incubate in a Thermocycler with the following program:

37°C for 50 minutes
70°C for 10 minutes
4°C forever

Once the temperature has reached 4°C, immediately add the following reagents:

1.1x µl DTT (100mM)
4.4x µl ATP (100mM)
5x µl Plasmid-Safe ATP-Dependent DNase (10U/µl)
3x µl Exonuclease I (20U/µl)

Incubate in a Thermocycler with the following program:

37°C for 30 minutes
80°C for 20 minutes
4°C forever

Purify the DNA by cold ethanol precipitation (-20 °C overnight) in the presence of 0.3M sodium acetate. After precipitation and wash (70% cold ethanol), dissolve DNA in a 50 µl Zymo Zyppy elution buffer. Since the majority of DNA will not be successfully circularized and will be digested by ATP dependent DNase and Exonuclease I, the DNA concentration at this point will be so low that only ethanol precipitation can lead to an acceptable recovery rate of DNA.

H. Fragmentation 2

Material: Covaris S220 Focused-ultrasonicator, Covaris microTUBE and genomic DNA.
Parameters for all libraries:

Peak intensity: 175
Duty cycle: 5%
Cycles per burst: 200
Treatment time: 45s
Volume: 50 µl
Temperature: 7°C

I. Immobilization with Streptavidin beads

Material: Dynabeads M-280 Streptavidin coated beads, 2X B&W buffer and fragmented DNA.
Immobilize DNA to the beads through the following steps:

Resuspend the beads and transfer 20 µl beads to a 0.5 ml PCR tube;
Remove the supernatant and wash the beads twice in 50 µl 2X B&W buffer;
Remove the supernatant, and add 50 µl 2X B&W buffer and 50 µl of fragmented DNA to the tube;
Incubate at 20°C for 1 hour on a rotator;
Remove the supernatant;
Wash the beads 4 times with 100 µl 2X B&W buffer and 2 times with 100 µl Zymo Zyppy elution buffer;
Remove the buffer and immediately proceed to the next step.

J. End repair 2

Material: NEBNext End Repair Module and immobilized DNA from last step.
Prepare the reaction in a 0.5 ml PCR tube:

All immobilized DNA from the last step
42.5 µl ddH_2O
5 µl End Repair Reaction Buffer (10X)
2.5 µl End Repair Enzyme

Incubate at 20°C for 30 min and keep at 4°C for 30 min. Remove the supernatant, wash 4 times with 100 µl 2X B&W buffer and 2 times with 100 µl Zymo Zyppy elution buffer. Remove the buffer and immediately proceed to the next step.

K. dA-tailing 2

Material: NEBNext dA-Tailing Module and end-repaired DNA.
Prepare the reaction in a 0.5 ml PCR tube:

All immobilized DNA from the last step
24.6 µl ddH_2O
3 µl dA-Tailing Reaction Buffer (10X)
2.4 µl Klenow fragment

Incubate at 37°C for 30 min. Proceed immediately to the next step without washing.

L. Ligation to TruSeq adapters

Material: NEBNext Quick Ligation Module, Illumina TruSeq adapters and dA-tailed DNA. Remember to use different adapters for different libraries if they will be sequenced on the same lane.

Prepare the reaction in a 0.5 ml PCR tube:

30 µl dA-tailing reaction mix from the last step
1 µl TruSeq adapter
4 µl ddH_2O
10 µl Quick Ligation Reaction Buffer (5X)
5 µl T4 Quick Ligase

Incubate at 20°C for 30 min. Remove the supernatant, wash 6 times with 100 µl 2X B&W buffer and 4 times with 100 µl Zymo Zyppy elution buffer. Remove the buffer and immediately proceed to the next step.

M. PCR amplification

Material: PCR Primer Cocktail, PCR Mater Mix from TruSeq DNA sample Prep V2 kit, and adapter-ligated DNA.

Prepare reaction in a 0.5 ml PCR tube:

All immobilized DNA from the last step
5 µl PCR primer cocktail
10 µl PCR master mix
35 µl ddH_2O

Carry out PCR in a thermal cycler with a heated-lid using the following program:

98°C for 30s
13 cycles of:
    98°C for 10s
    60°C for 30s
    72°C for 30s
72°C for 5 min
Hold at 4°C

Transfer the supernatant (the PCR products are in the solution, not on the beads) into another tube to perform purification with Ampure XP beads (1.1x volume) and elute in 20 µl Zymo Zyppy buffer to get the final library.

S1.4 Preparation for sequencing on the Illumina HiSeq platform

Measure the concentration of all the libraries by QPCR with the KAPA Library Quantification Kit for Illumina sequencing platforms following the manufacturer's protocol. We mixed 250 bp, 500 bp, 2 kb, 6 kb and 15 kb libraries to get the final library for sequencing with the relative molar concentration of each library being 40:20:8:4:3. The final library was sent to the genomics core facility at University of Texas Southwestern Medical Center to sequence 150 bp at both ends (PE150) with a rapid run on HiSeq1500.

S2 Genome assembly strategy

In the rest of this document, a line starting with $ indicates a command line used in that step to execute a certain program or script, and a line starting with * explains the command line. In the command lines, parameter values that should be changed for the specific cases are placed in square brackets "[]".

S2.1 Data processing and error correction

We obtained QSEQ format sequencing results from the genomics core facility, and processed them with in-house scripts to: (1) remove reads that did not pass the purity filter; (2) classify the reads according to the TruSeq adapter indices and (3) output the reads into FASTQ-format files. In addition, the mate pair libraries were processed by the Delox script[1] (loxP sequence: TCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACGT) to remove the loxP sequences (if any) from the reads and separate the true mate pair reads from paired-end reads. All sequence reads were then processed sequentially by the following procedures.

(1) mirabait from the MIRA package (MIRALIB version V3.4.0)[2] to remove reads contaminated by the TruSeq adapters and oligos (sequences stored in the file junk.fa) used in the sequencing reactions.

$ mirabait -ik 20 junk.fa [inputfile] [outputfile]

(2) fastq_quality_trimmer from the FASTX-Toolkits (V0.0.13)[3] to remove low quality (quality score < 20) portion at both ends and to discard reads shorter than 10 bp after trimming.

$ fastq_quality_trimmer -t 20 -l 10 -i [inputfile] -o [outputfile]

(3) JELLYFISH (version 1.1.2)[4] to obtain K-mer frequencies from reads in all the libraries.

$ jellyfish count -m [length of k-mer] -t 8 -s 10000000000 -c 8 --timing=jf.err --both-strands --min-quality=20 --stats=jf.stats [inputfiles]

$ jellyfish dump -ct mer_counts_0 > infiles.cts

$ cut -f 2 infiles.cts | sort -nrk 1 | uniq -c > [output_for_making_histogram]

The histogram of the 17-mer frequency is compared with a similar graph in the *Plutellas xylostella* genome paper, as shown in Fig. 1c of the main text. Another histogram of 19-mer frequency (supplementary Fig. S1a) shows that a cutoff of 7 times (frequency of a 19-mer) can separate the peak dominated by 19-mers with and without sequencing errors. This cutoff and the 19-mer counts from JELLYFISH were used subsequently to perform error correction with QUAKE[5]. We chose 19-mer frequency to determine this cutoff following the recommendation from QUAKE website (http://www.cbcb.umd.edu/software/quake/faq.html). They suggest to determine the length of k-mer based on k =log(200G)/log(4), where G is the size of the genome.

To run QUAKE (version 0.3), we prepared 3 files in the current directory: (1) cutoff.txt which has the cutoff for k-mer frequency; (2) "infiles" contains the paths for all the reads; (3) "infiles.cts" contains the JELLYFISH k-mer counts for reads in "infiles". The command for running QUAKE is:

$ quake.py -f infiles --int --no_cut --no_count -k [size of k-mer] -p [number_of_CPUs]

Afterwards, we used an in-house script find the corresponding pair for each read. Reads whose pairs were removed in previous steps were combined into a separate single-end library.

This data processing resulted in 9 libraries stored in 17 files that were used to assemble the genome: two paired-end libraries with insert sizes 250 bp and 500 bp, three paired-end libraries, three true mate pair libraries from the 2 kb, 6 kb and 15 kb libraries, and a single-end library containing all reads without pairs. JELLYFISH was applied a second time to the reads after error correction to generate a histogram for 17-mer frequencies.

S2.2 Estimation of the genome size

The histogram of 15-mer frequencies for all libraries (average read length 137 bp and total length 27,011,000,000 bp) after error correction (supplementary Fig. S1b) was used to estimate the genome size. This graph shows that 15-mers from homozygous regions are covered approximately 65 times. Therefore, the estimated coverage for the genome is k-mer coverage * average read length / (average read length – k-mer length + 1) = 65*137/(137-15+1) = 72.4 fold[6], and the estimated genome size is 27,011,000,000/72.4 = 373 million base pairs.

S2.3 Genome assembly

We tested three genome assemblers, including the well-established SOAPdenovo2 (version 2.04-r240)[7], ALLPATH-LG (version r43762)[8] and a new software designed for highly heterozygous genomes, Platanus (version 1.2.1)[9]. Both SOAPdenovo2 and ALLPATH-LG produced genome assembly with low scaffold N50 but a genome size much larger than the estimated size from K-mer. This indicates that the heterozygosity of the *Papilio glaucus* genome is too high for them to correctly distinguish equivalent regions in homologous chromosomes from duplication events. However, Platanus keeps track of the coverage for assembled regions in every stage of the assembling process and uses this information to detect and merge divergent equivalent regions in homologous chromosomes[9]. This strategy is suitable for many insect genomes with small size (indicates small number of gene duplication events) and high heterozygosity.

The performance of Platanus depends on the user-selected parameters, especially those defining the identity cutoff for merging divergent equivalent regions from homologous chromosomes. These parameters should depend on the heterozygosity level of a genome. For the *Papilio glaucus* genome with an overall heterozygosity rate of approximately 2% (see S5.1 for details), assembling with the following parameters produced an initial assembly with N50 comparable to published Lepidoptera genomes and assembly size (405Mbp) close to what we estimated based on K-mer (373Mbp). We call this initial result assembly_V0, which is the basis for the *Papiilo glaucus* genome draft. The commands used to produce assembly_V0 with Platanus are:

$ platanus assemble -f [fastq files for paired-end libraries] -t [number_of_CPUs] -o glaucus -m 128 -u 0.3 -d 0.3 -a 5

$ platanus scaffold -c glaucus_contig.fa -b glaucus_contigBubble.fa -IP1 [fastq files for 250bp paired-end library] -IP2 [fastq files for 500bp paired-end library] -IP3 [fastq files for paired-end libraries separated from mate pair libraries] -OP4 [fastq files for 2kb mate pair library] -OP5

[fastq files for 6kb mate pair library] -OP6 [fastq files for 15kb mate pair library] -u 0.3 -t [number_of _CPUs] -o glaucus

$ platanus gap_close -c glaucus_scaffold.fa -IP1 [fastq files for 250bp paired-end library] -IP2 [fastq files for 500bp paired-end library] -IP3 [fastq files for paired-end libraries separated from mate pair libraries] -OP4 [fastq files for 2kb mate pair library] -OP5 [fastq files for 6kb mate pair library] -OP6 [fastq files for 15kb mate pair library] -ed 0.1 -t [number_of_CPUs] -o glaucus.

S2.4 Post-assembly improvement

The reads from all libraries used by the assembler were mapped to assembly_V0 with Bowtie2[10] and the results were further processed by SAMtools[11]. This mapping allowed us to calculate the average coverage for each scaffold (supplementary Fig. S1c) and the number of reads mapped in each 100 bp sliding window (supplementary Fig. S1d). Both graphs illustrate that there are regions in the genome with about half of the expected coverage. As shown in supplementary Fig. S1c, the distribution of scaffold-level coverage has two peaks and the peak with higher coverage (center of this peak is about 67.5 fold coverage) corresponds to the expected coverage of a diploid genome.

We assumed that if we did not have highly heterozygous regions that were not merged together, the histogram of coverage from 0 to 135 (centered around 67.5) should be similar to a normal distribution with and the left shoulder decaying in a similar way as the right shoulder. As there are repeats in the genome, the left shoulder is expected to decay even faster. Base pairs with coverage from 68 to 83 accounts for 95% of all base pairs that fall into the right shoulder (coverage from 68 to 135). Therefore, coverage above 83 is significantly different (confidence level: 95%) from the expected coverage (67.5). If there were no heterozygosity problems (i.e., both shoulders are symmetric), the cutoffs for significantly lower coverage and significantly higher coverage should be centered around 67.5 fold as well. Therefore, we estimated the cutoff for significantly lower coverage as 67.5 * 2 – 83 = 52, suggesting that coverage less than 52 fold is significantly different (confidence level 95%) from the expected value for a diploid genome.

The scaffolds with low coverage were likely dominated by highly heterozygous regions that were not merged with the equivalent segments in homologous chromosomes. The presence of such segments also explains why the size of assembly_V0 is larger than expected. Therefore, scaffolds with coverage less than 52 were merged into other scaffolds if they could be nearly fully (coverage >90%, uncovered region < 500 bp) aligned to another low-coverage region in a longer scaffold with high sequence identity (>95%). For scaffolds with even lower coverage and smaller size (size < 1000 bp and coverage < 39 or size < 10000 bp and coverage < 20), we used a looser cutoff for identity (> 90%) to merge them into the longer scaffolds. The assembly after this step, namely assembly_V1, is the current genome assembly and is used for gene annotation and other analysis. The scripts used for this step are available at: http://prodata.swmed.edu/LepDB/.

Similar problems (larger than expected genome assembly size and regions with low coverage) occurred in the initial *Heliconius melpomene* genome assembly made by the CABOG assembler, because CABOG is not designed to work with heterozygous genomes[12]. The authors for *Heliconius* genome project adopted a strategy similar to ours to improve the initial assembly

and to remove the redundant scaffolds resulting from divergent equivalent regions from homologous chromosomes[13].

While the widely used genome assembler ALLPATH-LG discards all the scaffolds smaller than 1000 bp, Platanus keeps all scaffolds regardless of their length, resulting in a large number of scaffolds in genome assembly_V1 (68029). This number would be reduced to 20797 if all the scaffolds both shorter than 1000 bp and lacking annotated proteins were removed. However, we prefer to keep our genome as complete as possible, and thus we included all short scaffolds in the *Papilio glaucus* genome draft.

S3 Transcriptome assembly strategy

S3.1 Data processing

We used the reads from previously published RNA-seq libraries of 2 *Papilio glaucus*, 2 *Papilio canadensis*, 2 *Papilio appalachiensis* and 2 *Papilio polites* specimens[14]. These reads were downloaded from NCBI SRA database (Accessions: SRX277384, SRX277385, SRX277387, SRX277389, SRX277390, SRX277399, SRX277400, SRX277401). According to the previous publication, the *Papilio polytes* specimens came from a lab colony originating from the Philippines and RNA was extracted from their pupal wing discs. In contrast, the *Papilio glaucus*, *Papilio canadensis and Papilio appalachiensis* specimens were collected in the field from Louisiana, New Hampshire, and West Virginia, respectively, and RNA was extracted from the entire pupa[14]. These RNAseq libraries contain sufficient data for transcriptome assembly, ranging from 4.6Gbp to 15Gbp in each library. Similar to the procedure described in S2.1, reads with contamination from TruSeq adapters and the low quality portion of reads were removed using mirabait and fastq_quality_trimmer before they were supplied to the assemblers.

S3.2 *De novo* assembly, reference-guided assembly and mapping to the genome

We applied three methods to assemble the transcriptomes for *Papilio glaucus*, *Papilio canadensis* and *Papilio appalachiensis* specimens:

(1) *de novo* assembly by Trinity (version r20140413p1)[15,16]

$ Trinity --output .[output directory] --seqType fq --JM 100G --normalize_reads --left [RNAseq reads_1 in fastq format] --right [RNAseq reads_2 in fastq format] --CPU 24

(2) reference guided assembly by TopHat[17] (v2.0.10) and Cufflinks[18] (v2.2.1)

$ bowtie2-build [genome in fasta format] [indexed genome base name]

$ tophat --read-edit-dist 5 --fusion-read-mismatches 3 --segment-mismatches 3 --read-mismatches 4 --read-gap-length 4 --output-dir [output directory] --read-realign-edit-dist 0 --mate-inner-dist 100 --mate-std-dev 50 --solexa1.3-quals --num-threads 32 --coverage-search --b2-sensitive --library-type fr-unstranded [indexed genome base name] [RNAseq reads_1 in fastq format],[RNAseq_reads_2 in fastq format],[RNAseq single-end reads in fastq format]

$ cufflinks -p [number of CPUs] [TopHat alignments in bam format]

$ gffread -w transcripts.fa -o transcripts.gff -g [genome assembly in fasta format] transcripts.gtf

(3) reference guided assembly by Trinity based on TopHat's alignments.

$ Trinity –output [output directory] --normalize_reads --genome [genome assembly in fasta format] --genome_guided_max_intron 100000 --genome_guided_sort_buffer 18G --seqType fq --JM 18G --genome_guided_use_bam [TopHat alignment in bam format] --left [RNAseq reads_1 in fastq format] --right [RNAseq reads_2 in fastq format] --CPU 6 --genome_guided_CPU 6 --GMAP_CPU 6

The results from all three methods were then integrated by Program to Assemble Spliced Alignments (PASA, version r20130907)[19,20] with the following commands:

$ cat [Trinity *de novo* assembly in fasta format] [Trinity genome guided assembly in fasta format] > [Trinity assemblies]
$ seqclean [Trinity assemblies] -c 8
$ accession_extractor.pl < [Trinity assemblies] > tdn.accs
$ Launch_PASA_pipeline.pl -c alignAssembly.config -C -R -g [genome assembly in fasta format] -t [Trinity assemblies after seqclean] -T -u [Trinity assemblies before seqclean] --TDN tdn.accs --cufflinks_gtf [Cufflinks result in gtf format] --ALIGNERS blat,gmap --CPU [number of CPUs]
PASA also mapped the transcripts to the *Papilio glaucus* reference genome. This mapping was used not only as the basis for identifying orthologs between species, but also for transcript-based gene annotation. For the more distantly related *Papilio polytes*, only *de novo* assembly by Trinity was performed.

S4 Genome assembly quality assessment

We obtained the most recent versions of all published Lepidoptera genomes, including *Bombyx mori*, *Danaus plexippus*, *Heliconius melpomene* and *Plutella xylostella*[13,21-27], and compared their quality to the *Papilio glaucus* genome. In addition to the continuity reflected by N50, completeness is another very important indicator of genome quality. We evaluated the completeness of these genomes by analyzing the coverage of independently obtained transcripts, Core Eukaryotic Genes Mapping Approach (CEGMA)[28] genes and the Cytoplasmic Ribosomal Proteins. The evaluation was done using the criteria that were used in the Monarch butterfly genome paper[26].

S4.1 Genome assembly quality assessment by the coverage of transcripts

We adopted the criterion used in the Monarch butterfly genome paper, and considered a transcript to be covered if the e-value of its best BLASTN[29] hit in the genome of the same species is smaller than 10^{-50}. The *de novo* assembled transcriptomes from two *Papilio glaucus* specimens were used to evaluate the completeness of the *Papilio glaucus* genome. 97.8% (63,592 out of a total 65,018) and 98.0% (47,204 out of 48,177) of transcripts from them meet the criterion, respectively. The number of transcripts assembled by Trinity is large, as many of them are redundant with several transcripts mapping to the same loci. Similar statistics for *Danaus plexippus* and *Bombyx mori* were taken from the Monarch butterfly genome paper.

The transcriptomes from several samples of *Plutella xylostella* were downloaded from the Diamondback moth Genome Database (DBM-DB)[25]. Of these 171,262 transcripts, a surprisingly large portion (29,260, 17.1%) does not meet the criterion above (cannot find hit in the *Plutella xyostella* genome with e-value less than 10^{-50}). Obviously, transcripts that are truly covered by the genome could have an e-value above 10^{-50}, due to being composed of only short exons or due to the high variation rates (correlates well with heterozygosity) between *Plutella xylostella* specimens. However, since the same criterion is used for *Papilio glaucus,* which has a comparable level of heterozygosity (Fig. 1c), the significantly reduced number of transcripts that map confidently to the reference genome instead reflects the relatively poor completeness of the *Plutella xylostella* genome. This is supported by another, independent test aiming to identify Hox genes, in which 2 out of the 14 conserved Hox genes are missing in the *Plutella xylostella* genome, but are present in its transcriptomes. This also allows us to estimate that about 14% genes are missing in the *Plutella xylostell*a genome draft. Apparently the innovative approach used for that genome in order to overcome the problems caused by high heterozygosity (i.e. dividing the genome into fosmid clones to sequence and assemble separately, followed by merging these fragments together to obtain the whole genome) missed a considerable portion of the genome.

S4.2 Genome assembly quality assessment by the coverage of CEGMA genes

The 457 core eukaryotic genes (CEGMA genes) from *Drosophila melanogaster* were used to evaluate the completeness of Lepidoptera genomes and a gene was considered to be covered by the genome if its best TBLASTN hit in the genome had an e-value lower than 10^{-5}. By this criterion, three CEGMA genes (0.7%) are missing in the *Papilio glaucus* genome. However,

the orthologs of these three genes are consistently missing in all independently sequenced Lepidoptera genomes. Therefore, possibilities other than genome incompleteness are more likely responsible for their absence: (1) their sequences in Lepidoptera genomes diverged a lot from *Drosophila*; (2) they are made of short exons in Lepidoptera genomes; and (3) they are not essential and lost in Lepidoptera.

To test whether the scaffolds in the genome assemblies are long enough to completely cover most of the protein coding genes, for each CEGMA gene we calculated the percentage of residues that were covered by the most confident scaffold in the TBLASTN alignment. Judging by this criterion, *Papilio glaucus* is among the best (only worse than the *Danaus plexippus* genome) with an overall coverage of 85.6% (average coverage weighted by the length of each genes). The modest coverage at the residue level is expected, due to the presence of short exons and the distant relationship between *Drosophila* and Lepidoptera.

S4.3 Genome assembly quality assessment by Cytoplasmic Ribosomal Proteins

We searched Flybase[30,31] with the term "ribosomal proteins" and selected 93 Cytoplasmic Ribosomal Proteins manually from the result. We considered a Cytoplasmic Ribosomal Protein to be present in a Lepidoptera genome if its best TBLASTN hit in the genome had an e-value below 10^{-5}. 100% of these CRPs are present only in the *Papilio glaucus* and *Danaus plexippus* genomes, again placing it among the best genomes.

S4.4 Genome assembly quality assessment by consistency with the paired-end reads

We applied REAPR (Recognising Errors in Assemblies using Paired Reads, version 1.0.17) to evaluate the quality of the genome assembly. REAPR maps the reads to the genome assembly and calls regions with a significant drop in coverage comparing to neighboring regions, they likely corresponds to regions where assembly error occurs. We used the following commands to execute this assessment:

```
$ reapr perfectmap assembly.fa short_1.fq short_2.fq 260 perfect
$ reapr smaltmap -n 30 assembly.fa long_1.fq long_2.fq long_mapped.bam
$ reapr pipeline assembly.fa long_mapped.bam output perfect
```

The major estimates we got from the program were: Error free bases: 89.54%; FCD errors within a contig: 213 (out of 101904 contigs, that is 0.2%); and FCD errors over a gap: 546 (count of 33875 gap, that is 1.6%). These statistics are rather good compared to genomes that were used to benchmark REAPR. For example, for human genome (GRCh37), REAPR also reports FCD errors over 1.7% of gaps and suggests only 79.1% of the base pairs in the genome assembly are error-free. REAPR also tries to break the assembly at places where an assembly error may occur. As a result, it generated another error-free genome assembly with scaffold N50 of 186 kb.

S5 Detection of SNPs in the genome

S5.1 SNP detection and overall SNP rate

The reads from all the libraries used to assemble the genome were mapped to the *Papilio glaucus genome* with Bowtie2 (v2.1.0). We applied Bowtie2 to libraries with different insert sizes separately as they requires different values for the "--maxins" parameter, which defines the maximal insert size of that library. For 250bp, 500bp, 2kb, 6kb, and 15kb libraries, we used 500, 1200, 5000, 10000 and 20000 for this parameter, respectively.

$ bowtie2-build [genome in fasta format] [indexed genome base name]

$ bowtie2 --threads 32 --un [unpaired report] --un-conc [discordant report] --maxins [value depends on the library size] --phred64 --end-to-end --sensitive -x [indexed genome base name] -1 [library reads_1] -2 [library reads_2] -S [output in SAM format]

This mapping allowed us to identify SNPs, insertions and deletions that happened to only one of the homologous chromosomes. We implemented two methods for this task:

(1) the SAMtools (version 0.1.19)/BCFtools (version 0.1.19)[11] pipeline.

* combine the SAM format output from Bowtie2 for all the libraries into one file: [Bowtie2 alignments in SAM format]

$ samtools view -bS [Bowtie2 alignments in SAM format] > [Bowtie2 alignments in BAM format]

$ samtools sort [Bowtie2 alignments in BAM format] [sorted Bowtie2 alignments in BAM format]

$ samtools index [sorted Bowtie2 alignments in BAM format]

$ samtools faidx [genome assembly in fasta format]

$ samtools mpileup -g -f [genome assembly in fasta format] [sorted Bowtie2 alignments in BAM format] > [input file for bcftools]

$ bcftools view -bvcg [input file for bcftools] > [bcftools output in bcf format]

$ bcftools view [bcftools output in bcf format]> [bcftools output in vcf format]

$ vcfutils.pl varFilter -D200 [bcftools output in vcf format] > [filtered SNP calls in vcf format]

(2) the Genome Analysis Toolkit[32,33] (GATK).

$ java -jar CreateSequenceDictionary.jar REFERENCE=[genome assembly in fasta format] OUTPUT=[genome assembly as a dictionary]

$ java -jar SortSam.jar INPUT=[Bowtie2 alignments in SAM format] OUTPUT=[step1 BAM format output] SORT_ORDER=coordinate

$ java -jar MarkDuplicates.jar INPUT=[step1 BAM format output] OUTPUT=[step2 BAM format output] METRICS_FILE=metrics.txt

$ java -jar AddOrReplaceReadGroups.jar I=[step2 BAM format output] O=[step3 BAM format output] SORT_ORDER=coordinate RGID=group1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=[genome assembly base name] CREATE_INDEX=True

$ java -jar BuildBamIndex.jar INPUT=[step3 BAM format output]

$ java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -fixMisencodedQuals -R [genome assembly in fasta format] -I [step3 BAM format output] -o target_intervals.list

$ java -jar GenomeAnalysisTK.jar -T IndelRealigner -fixMisencodedQuals -R [genome assembly in fasta format] -I [step3 BAM format output] -targetIntervals target_intervals.list -o [step4 BAM format output]
$ java -jar GenomeAnalysisTK.jar -fixMisencodedQuals -l INFO -R [genome assembly in fasta format] -T UnifiedGenotyper -I [step4 BAM format output] -o [SNP calls in VCF format] --output_mode EMIT_ALL_SITES

According to GATK, the mapped reads offered sufficient statistics to detect SNPs for 336,752,378 positions, and SNPs were detected at 6,750,896 positions, implying a heterozygosity level of 2.0%. The SAMtools/BCFtools pipeline detected fewer (5,050,690) SNPs. 4,314,123 SNPs were detected by both methods, reflecting a good overlap (85.4%) between the results of the two methods. The 5,050,690 SNPs detected by the SAMtools/BCFtools pipeline was used in the following analysis.

S5.2 Distribution of SNPs in different genomic regions

To analyze the distribution of SNPs, we divided the genome into different regions, i.e. exons, introns, repeats and intergenic regions. Two related indicators, i.e. the percentage of SNPs in 1000 bp windows and the distances between neighboring SNPs, were used to reflect this distribution. Based on the average SNP rate, we simulated a random distribution of SNPs in the genome, assuming equal probability for a SNP to happen to any base pair (grey lines in Fig. 1d of main text and supplementary Fig. S1e). The real distribution of SNPs is non-random and SNPs tend to concentrate in certain regions. The SNP rate in exons is generally much lower than in non-coding regions, but there are still exons with a high density (> 3%) of SNPs.

S5.3 Proteins enriched in SNPs

We identified proteins with significantly more SNPs by means of binomial tests (p = average SNP rate in all exons, m = number of SNPs in the exons of a protein, N = length of the exons of a protein). To avoid false discoveries simply due to the large number of statistical tests performed, we carried out False Discovery Rate tests and calculated the Q-values[34]. We consider proteins with Q-values smaller than 0.1 to be significantly enriched in SNPs.

The GO terms[35] and their parental GO terms associated with these SNP-enriched proteins were extracted, counted and compared with a background of GO terms (and their parental GO terms) associated with all *Papilio glaucus* proteins. Significantly enriched GO terms were selected by means of binomial tests (p = probability for the GO term to be associated with any protein in the background, m = number of SNP-enriched proteins associated with this GO term, N = number of proteins with significantly enriched SNPs). The significantly enriched GO terms (P-value < 0.01) were submitted to the REVIGO web server to cluster GO terms by similarity in meaning and to visualize them.

S6 Identification and classification of repeats

S6.1 Construction of a species-specific repeat library

To annotate the repeats and transposable elements in the *Papilio glaucus* genomes, we first used the RepeatModeler (version 3.0.9)[36] pipeline to detect species-specific repeat families. This pipeline employs two *de novo* repeat predictors, RECON[37] and RepeatScout[38] to identify similar sequences that repeatedly appear in different loci of the genome. We used the following commands and the final product of RepeatModeler is a list of representative sequences of repeat families in the genome.

$ BuildDatabase -name [database name] [genome assembly base name]
$ RepeatModeler -database [database name]

In addition, mapping reads to the genome reveals regions with significantly high coverage. This is true for genomes assembled with different methods. It is likely that such regions correspond to repeats in the genome that have not yet diverged. Due to the high sequence similarity (approaching 100%), genome assemblers merge them. We used in-house scripts to identify these repeats based on the number of reads mapped in 100 bp windows in the genome (introduced in S2.3). As shown in supplementary Fig. S1d, the histogram of the number of reads mapped to homozygous regions has a peak at 124 reads. We considered any 100 bp windows with more than 620 mapped reads to be from repeats and we joined neighboring 100 bp windows satisfying this criterion to obtain the complete repeat sequences.

The repeat sequences detected by the two methods above were submitted to the CENSOR[39] web server (http://www.girinst.org/censor/) to assign them to the repeat and transposable element classification hierarchy. These repeats, and their classification status, were included to construct a species-specific repeat library.

S6.2 Detection and masking of repeats in the genome

The species-specific repeat library and the repeats classified in RepBase[40] (V18.12) were used to identify and mask repeats in the *Papilio glaucus* genome by RepeatMasker (version 3.0.9)[41] with the following commands:

$ RepeatMasker -lib [species specific repeat library] -pa 32 -div 30 [genome assembly in fasta format]
$ RepeatMasker -species all -pa 32 -div 30 [genome assembly with repeats masked from the last step]

We used a diversity cutoff of 30% (-div 0.3) for RepeatMasker, and detected 388,266 simple and interspersed repeats that comprise 22.2% of the genome (83.4 Mbp). This value is comparable to those for butterflies, but lower than those for moths. However, the number of repeats one can identify is sensitive to the procedures and parameters used in the data analysis, such as the cutoff of sequence divergence level set in RepeatMasker. Therefore, it is difficult to conclude whether the difference in the repeat content of these genomes is significant unless the repeats are identified with the same procedure.

S7 Gene annotation

We annotated the protein coding genes with a pipeline very similar to what is implemented in the Broad Institute[42]. In short, transcript-based, homology-based and *de novo* approaches were used to generate 15 different sets of gene annotations. These annotations were combined with EvidenceModeller (version r2012-06-25)[20] to generate consensus-based final predictions.

S7.1 Transcript-based gene annotation

As described in S3.2, we assembled the transcriptomes of two *Papilio glaucus* specimens using different pipelines. For each specimen, we obtained two sets of transcript-based annotations from (1) a pipeline containing TopHat and Cufflinks and (2) a more sophisticated pipeline that uses PASA to integrate the results from Trinity *de novo* assembly, Trinity reference-guided assembly and the result from TopHat and Cufflinks. In total, these approaches produced 4 sets of gene annotations.

S7.2 Homology-based gene annotation

The protein sets from all four published Lepidoptera genomes, and the *Drosophila melanogaster*[43] in Flybase, were used as references to annotate *Papilio glaucus* proteins with the exonerate (version 2.2.0) software[44]. For each reference protein, we used the following command to produce a homology-based gene annotation.

$ exonerate --model protein2genome --refine region –q [reference protein sequence in fasta format] -t [genome assembly in fasta format] -Q protein -T dna --showtargetgff yes --showalignment yes --percent 30 > [output]

We enforced an identity cutoff of 30% to reduce the number of imperfect gene models based on remote homologs. This approach produced 5 sets of gene annotations that are based on different reference organisms.

In addition, proteins from the entire UniRef90[45] (Mar. 2014) database were used as references to annotate genes. For this large data set, we used genblastG (version 1.39)[46], a new, faster and splicing-site aware software that is similar to and is claimed to work no worse than exonerate. For each reference protein, we used the following command:

$ genblast -p genblastg -q [reference protein sequence in fasta format] -t [genome assembly in fasta format] -g T -v 2 -c 0.5 -e 0.00001 -s 0 -o [output base name] -gff -cdna -pro

The parameters "-g T -v 2 -c 0.5 -e 0.00001 -s 0" were specified for genblastG to limit the number of gene models with poor support. This approach generated 227,887 redundant gene models (several of them could map to the same loci) and all of them were used as one set of gene annotations.

S7.3 *De novo* gene annotation

One essential step for *de novo* gene annotation is to train the predictors with confident gene annotations. We manually curated and selected 1313 confident gene models by integrating the evidence from transcripts and homologs with the help of in-house scripts. For

homology-based predictions, only those models based on *Drosophila melanogaster*, *Danaus plexippus* and *Bombyx mori* proteins were used, because much effort has been made on the annotation proteins in these species. A confident gene model in *Papilio glaucus* needs to satisfy the following criteria: (1) both the homology-based methods and the transcript-based methods consistently predict the splicing sites inside this gene; (2) the predicted gene is completely covered by a transcript; (3) the predicted gene has a standard translation initiation site and stop codon.

We implemented four *de novo* gene predictors, AUGUSTUS (version 2.6.1)[47], SNAP (version 2006-07-28)[48], Genemark (version 2.3c)[49] and GlimmerHMM (version 3.0.1)[50]. Genemark is able to train itself on the input whole genome data with the following command:
$ gm_es.pl --max_nnn 1000 [input genome assembly in fasta format]

Other gene predictors were trained with our manually selected good gene models following the instructions from each program.
For AUGUSTUS, we used the following commands (scripts are from AUGUSTUS package):
$ perl gff2gbSmallDNA.pl [curated gene models in GFF format] [genome assembly in fasta format] 1000 [training set in Genbank format]
$ perl new_species.pl --species=[species name]
$ perl optimize_augustus.pl --cpus=32 --species=[species name] [curated gene models in Genbank format]
$ etraining --species=Papilio_glaucus [curated gene models in Genbank format]

For SNAP, we used the following commands (linux commands or scripts from SNAP package):
$ perl gff2zff.pl < [curated gene model in GFF format] > [training set in ZFF format]
* prepare the scaffold sequences in the same order as they show up in file [training set in ZFF format] and store them in file [sequences for training]
$ fathom [training set in ZFF format] [sequences for training] -gene-stats
$ fathom [training set in ZFF format] [sequences for training] -validate
$ fathom [training set in ZFF format] [sequences for training] -categorize 1000
$ fathom uni.ann uni.dna -export 1000 -plus
$ mkdir params
$ cd params/
$ forge ../export.ann ../export.dna
$ cd ..
$ hmm-assembler.pl [species name] params/ > [SNAP trained parameters]

For GlimmerHMM, we used the following commands:
$ python zff2glim.py [training set in ZFF format] > [training set for GlimmerHMM]
$ trainGlimmerHMM [sequences for training] [training set for GlimmerHMM] -b 2

For GlimmerHMM and Genemark, we used the genome sequence without repeat masking as input, because they do not handle masked repeats properly. Maker[51], a widely used gene annotation pipeline, also uses a genome sequence without masking as an input to Genemark. AUGUSTUS and SNAP's performance can be significantly improved if evidence-

based (transcripts and homologs) gene predictions are supplied to them. We used the Maker[51] pipeline to obtain evidence-guided predictions from AUGUSTUS and SNAP and *de novo* predictions from Genemark. In addition, we supplied all evidence-based and *de novo* predictions to Maker, so that it could make consensus-based predictions. Predictions made by Maker are similar to those predicted by *de novo* predictors and we retained them as an additional set of *ab initio* predictions. Thus, we constructed 5 sets of gene annotations made by *de novo* predictors.

Although Maker, AUGUSTUS and SNAP use homology and transcript-based evidence to assist gene prediction, we still consider their predictions to be *de novo*, because all these programs consider intrinsic features of the genomic sequence, such as quality of the open reading frames and the presence of transcription and translation initiation sites, to make their predictions. Consideration of these intrinsic features is the essence of *de novo* gene prediction.

S7.4 Consensus-based final gene annotation

All 15 sets of gene predictions discussed above and the annotation of repeats were integrated by EvidenceModeller to make the final gene predictions. As recommended by the author, we weighted transcript-based predictions more than homology-based ones, and weighted *de novo* predictions the least. For predictions that tend to be more reliable, a higher weight was given[51,52]. The weights we assigned for all the annotation resources are:

```
PROTEIN         exonerate:Hm 3
PROTEIN         exonerate:Dp  10
PROTEIN         exonerate:Px   2
PROTEIN         exonerate:Dm 5
PROTEIN         exonerate:Bm 5
PROTEIN         genBlastG       5
TRANSCRIPT    cufflinks:A       5
TRANSCRIPT     cufflinks:B     5
TRANSCRIPT    pasa:A 10
TRANSCRIPT     pasa:A  10
ABINITIO_PREDICTION     maker         10
ABINITIO_PREDICTION           augustus         4
ABINITIO_PREDICTION           snap     3
ABINITIO_PREDICTION           genemark         2
ABINITIO_PREDICTION           GlimmerHMM 1
```

We used the following commands to get EvidenceModeller predictions:

*all repeats annotation in file “repeats.gff3”, all the transcript-based annotations in file “transcript.gff3”, all the homology-based annotations in file “homolog.gff3” and all the *de novo* prediction in “denovo.gff3”

$ perl partition_EVM_inputs.pl --genome [genome assembly in fasta format] --gene_predictions denovo.gff3 --protein_alignments homolog.gff3 --transcript_alignments transcript.gff3 --repeats repeats.gff3 --segmentSize 2000000 --overlapSize 10000 --partition_listing partitions_list.out

$ perl write_EVM_commands.pl --genome [genome assembly in fasta format] --gene_predictions denovo.gff3 --protein_alignments homolog.gff3 --transcript_alignments transcript.gff3 --repeats repeats.gff3 --output_file_name evm.out --partitions partitions_list.out --weights [file with weights listed above] --search_long_introns 1 --re_search_intergenic 1 > commands.list
* carry out all the commands in "commands.list" on multiple CPUs.

For proteins and the protein families that were further analyzed in the manuscript, we manually curated the gene models from EvidenceModeller and modified a small number of gene models. This manual curation resulted in the detection of a few genes missed by EvidenceModeller, which we added to the final gene set.

S7.5 Prediction of protein function and additional features

The well-curated protein annotations in the Swissprot[53] database have been shown to be of high quality[54]. Therefore, we predicted the function of *Papilio glaucus* proteins by transferring annotations from the closest BLAST hit in Swissprot, requiring the e-value to be less than 10^{-5}. This approach annotated 11065 proteins. In addition, for each protein, we identified its closest *Drosophila melanogaster* homolog in Flybase and detected confident homologs (e-value < 10^{-5}) for 11,816 proteins. This mapping provides a better description of the putative function for each protein by linking the rich information and literature associated with the *Drosophila* protein to the *Papilio glaucus* protein. The GO terms associated with the closest *Drosophila* homologs were transferred to the *Papilio glaucus* proteins, and thus we obtained associated GO term annotations for 10805 proteins. Combining both approaches, we were able to predict the functions of 11975 *Papilio glaucus* proteins.

Finally, we applied the comprehensive pipeline, InterproScan (version 5.6)[55], to every *Papilio glaucus* protein to identify conserved protein domains[56-62] and functional motifs[60,63], to predict sequence features including coiled coil[64], transmembrane helices[65,66] and signal peptides[66,67], to detect homologous structures[68,69] that can be used for structure prediction, to assign *Papilio glaucus* proteins to protein families[56-62] and to map them to metabolic pathways[70-73]. For each protein, we ran InterproScan with the following command:
$ interproscan.sh -i [protein sequence in fasta format] -b result/ -dp -goterms -pa

S8 Comparison of Lepidoptera genomes

S8.1 Identification of orthologs and evolutionary analysis

We compared the *Papilio glaucus* protein set with the official protein sets from all published Lepidoptera genomes, including *Bombyx mori*, *Danaus plexippus*, *Heliconius melpomene* and *Plutella xylostella*. We used OrthoMCL (version 2.0.9)[74] to identify the orthologous protein groups from these species. We followed the User Guide came with the OrthoMCL package. Briefly, after modifying the configure file "orthomcl.config" to indicate database names and login information for MySQL, the following commands were used:

$ orthomclInstallSchema my orthomcl.config install_schema.log

* for each species, combine the proteins sequences in one fasta format file. For each fasta file, do the following command:

$ orthomclAdjustFasta [species name abbreviation] [protein sequence in fasta format] [id_field]

* this command will produce input fasta files for the next step in "./compliantFasta" directory.

$ orthomclFilterFasta ./compliantFasta/ 10 20

* this step will produce filtered sequences from all species in goodProteins.fasta

* for all protein sequences in goodProteins.fasta, do All-against-All BLASTP comparison and save the results in goodProteins.blast

$ orthomclBlastParser goodProteins.blast ./compliantFasta >> similarSequences.txt

$ orthomclLoadBlast orthomcl.config similarSequences.txt

$ orthomclPairs orthomcl.config

$ orthomclDumpPairsFile orthomcl.config

$ mcl mclInput --abc -I 1.5 -o mclOutput

$ orthomclMclToGroups [prefix of group names] [starting number for group names] < mclOutput > groups.txt

Groups made of single-copy orthologs from all genomes were extracted from OrthoMCL output and used to build a phylogenetic tree: (1) proteins from each selected group were aligned with MAFFT[75] using the following command and the aligned positions with any gaps were removed:

$ mafft --maxiterate 1000 --genafpair in > out

(2) the resulting gap-free alignments were concatenated to get the final alignment that was used to build an evolutionary tree with PHYML 3.0[76,77] (JTT model[78]); (3) this tree was visualized in FigTree[79].

S8.2 Synteny

By means of in-house scripts (available at http://prodata.swmed.edu/LepDB/), we calculated the percentage of proteins that are in micro-syntenic blocks between all pairs of Lepidoptera species. We used two criteria to determine whether a protein is in a micro-syntenic block shared between two species: the loose criterion requires one of its closest neighbors to be from the same orthologous group and in the same orientation, while the stringent criterion requires both closest neighbors to be from the same orthologous group and in the same orientation. We ignored neighbors from the same orthologous group of a gene because a

recent gene duplication event of one gene will not likely destroy the close localization and linkage between genes. To avoid influence from the length of scaffolds, only proteins with neighbors from different orthologous groups on both sides and in both species were considered, and we considered others to be without sufficient evidence.

We also identified gene inversion events, where two neighboring genes remained adjacent but reversed their order in the genome. The frequency of such rare events could be a good indicator of evolutionary distance[80] between species and was used to build a synteny-based evolutionary tree using the BioNJ[81] method from the phylogeny.fr web server[82].

S8.3 Analysis of Hox genes

Starting with all homeodomains from *Drosophila* in the HomeoDB[83], we identified all homeodomains in Lepidoptera genomes using BLASTP (e-value cutoff 0.001). We made a multiple sequence alignment of all the Lepidoptera homeodomains with Muscle[84]. Muscle attempts to cluster similar sequences together, and on the basis of this clustering, we manually clustered these homeodomains into orthologous groups. The clustering was trivial because homeodomain sequences in the same orthologous groups are frequently almost identical, except for 15 divergent ones that are mapped to the loci corresponding to the *Drosophila* Hox genes, *Zen* and *Zen2*.

We then focused on the homeodomains from the Hox genes. We identified Hox genes from unpublished, draft genomes for *Manduca sexta*[85] and *Helicoverpa punctigera*[86] as well. We built an evolutionary tree for homeodomains in Hox genes with Phyml (JTT model) implemented in the phylogeny.fr web server. This tree was visualized in FigTree. We mapped the homeodomains from Hox genes to the genomes. Their order in different genomes is mostly conserved. Even for the 3 *Zen*-like Hox genes that were difficult to classify, it is mostly true that homeodomains that are mapped to equivalent positions in the genome are grouped in one clade in the tree (except for *Bombyx mori*, which could be an artifact of the tree-building algorithm). The expansion of *Zen*-like genes is a common feature of all the Lepidoptera species[87] that we analyzed and there is an additional, significant expansion of *Zen*-like genes in *Bombyx mori*. However, their poor conservation compared to other Hox genes suggests that they may not play an important role and might even be pseudogenes.

S8.4 Identification of expanded gene families

We classified the Lepidoptera proteins in whole genomes into families on the basis of orthologous groups identified by OrthoMCL and the mapping of these proteins to the *Drosophila melanogaster* proteins in FlyBase by BLAST (e-value < 10^{-5}). If two OrthoMCL-defined orthologous groups overlapped in the *Drosophila* proteins to which they map, we merged them into a single protein family. This approach allowed us to group most proteins with the same function or highly similar functions together.

Two criteria were used to identity expanded gene families in *Papilio glaucus*: (1) *Papilio glaucus* must have more proteins from this family than from any other Lepidoptera species; (2) the total length of *Papilio glaucus* proteins in this family must be at least 1.5 times the total protein length for any other Lepidoptera. Proteins satisfying both criteria are listed in Table S15

and ranked by the minimum of two gene expansion indices: (1) the ratio of *Papilio glaucus* protein number to the average protein number in other species and (2) the ratio of *Papilio glaucus* protein total length to the average total length for other species.

S8.5 In-depth study of important gene expansion events

The most interesting and most confident gene expansion events were further investigated. For each family, the following steps were taken to ensure the inclusion of all relevant proteins: (1) search for homologs (e-value < 10^{-5}) in all Lepidoptera protein sets starting with all current members in a family; (2) annotate all hits by transferring the annotation from the best BLAST hit (e-value < 10^{-10}) in the Swissprot database; (3) remove the proteins that are remotely related and of different function based on BLAST statistics and function annotation (for example, other G protein coupled receptors that are not Opsins are removed from the Opsin family); (4) use proteins remaining after step 3 to search against the genome sequences by genblastG to obtain additional proteins that were missed in the official gene sets; (5) group proteins from the previous two steps according to their genomic loci and at each loci, select the best (by length and similarity to other proteins) gene model and remove other redundant models. Usually, we preferred to select gene models that are the same as those in the official protein sets since such models might be supported by RNA sequences. However, we sometimes modified these gene models, e.g. by extending the coding sequence or by separating a fused protein into two, so that they became more consistent with orthologs in other species. All these steps were performed with in-house scripts combined with manual curation.

One problem with interpretation of highly heterozygous genome is that highly divergent alleles from homologous chromosome may still appear in the genome assembly as two segments and they can be misinterpreted as duplication. We confirmed that was not the case for the expanded gene families we studied here using the following two criteria: (1) the sequence identity between a pair of proteins should be below 95%. (heterozygosity level at the DNA level for coding region is about 0.8%); (2) the coverage of the coding genes by the sequence reads should be about the expected value for a diploid genome.

Protein sequences from each protein family were aligned with MAFFT. However, for the homologs of farnesyl pyrophosphate synthase (FPS), we used Promals3D[88], a more advanced multiple sequence aligner that uses sequence and secondary structure profiles, to align them with homologs (human and chicken farnesyl pyrophosphate synthase) that have available three dimensional structures. Evolutionary trees were then built using PhyML (JTT model) on the phylogeny.fr web server and visualized in FigTree. The 3D structure templates for farnesyl pyrophosphate synthase homologs (PDB ID: 1FPS and 4GA3) were analyzed in Pymol to find residues relevant to catalytic activity and ligand binding.

To investigate whether the farnesyl pyrophosphate synthase family underwent expansion in other Papilionidae species, we used TBLASTN to search against the *de novo* assembled transcriptomes of *Papilio canadensis*, *Papilio appalachiensis* and *Papilio polytes*[14]. The results were manually checked and the numbers of non-redundant hits from these species were counted.

S9 Investigation of speciation between *Papilio glaucus* and *Papilio canadensis*

S9.1 Identification and alignment of orthologs

As described in S3.2, we used PASA to map the transcripts from *Papilio glaucus* and *Papilio canadensis* specimens to the reference genome. Transcripts from different specimen that are mapped to the same genomic loci in the reference genome are considered to be orthologous. PASA reports the mapping range and the corresponding transcript predicted in the reference genome, but not the alignments between them. Therefore, we used the accurate mode of MAFFT (--maxiterate 1000 –genafpair) to align the orthologous transcripts. Due to the high similarity between these transcripts, MAFFT alignments are usually accurate with the exception of the occasional error in aligning a shorter sequence to a longer one. For such cases, we performed local sequence alignments with BLAST as well and took the consistently aligned positions from both local and global alignments.

We selected regions shared by all specimens from each alignment and an orthologous group was included in the following analysis if this shared region was longer than 50 bp. In total, the 8230 protein coding genes that satisfy this criterion were used in all the analyses described below. We calculated the mutation rate for each gene and the overall mutation rate between any two specimens.

S9.2 Isolation with migration model

A meaningful reconstruction of the "Isolation with Migration" model by IMa2 (version August 27, 2012)[89] requires the input loci to be free from recombination. Proteins that showed significant signs of recombination in either the PHI permutation test or the Chi^2 test implemented in PhiPack (no version information available)[90] were removed from the analysis with the following command:

```
$ Phi -o TRUE -f [protein sequence in fasta format]
```

The remaining 6843 proteins were randomly divided into 20 data sets and IMa2 was applied to each of them separately.

For each IMa2 run, the first 200000 Monte Carlo steps (burn-in) were discarded to remove the influence of initial random parameters and to allow the simulated parameters to reach plateau. Starting with the parameters after burn-in, Monte Carlo simulations were carried out on multiple CPUs in parallel. The command for Burn-in was as below:

```
$ IMa2 -i [IMa2 input file with sequence alignment] -o [output base name] -s [random seed] -b10000 -t10.0 -m0.4 -q25.0 -l540 -hfg -hn100 -ha0.99 -hb0.75 -r2345 -z100 -p35 -u0.5 -hfg
```

The command for actual simulation was:

```
$ IMa2 -i [IMa2 input file with sequence alignment] -f [saved Markov chain state file after burn-in] -s [random seed] -b10000 -t10.0 -m0.4 -q25.0 -l540 -hfg -hn100 -ha0.99 -hb0.75 -r2345 -z100 -p35 -u0.5 -hfg
```

We stopped the simulation after about 3,500,000 Monte Carlo steps were made, combining all parallel threads and 35,000 steps were saved (save every 100 steps). The results

from multiple threads (*.ti files) were combined and analyzed using IMa2 L mode to generate the final predictions with a command like this:
$ IMa2 -i -vmrun –o [output result] -c2 -b10000 -t10.0 -m0.4 -q25.0 -l540 -r0 -y1 -p35 -u0.5
The results from IMa2 were interpreted and visualized using IMfig (no version information available)[91] with the following command:
$ python IMfig.py –i[IMa2 result] –o[output figure] –c

The results are summarized in Table S20. The estimated effective population sizes, species time and migration rates from the IMa2 simulations on the 20 data sets were averaged to obtain the final result. While the isolation time (0.51±0.03 million years) and effective population sizes of the ancestral species (419±29 thousands) estimated on different data sets agree with each other very well, the estimated migration rates vary considerably between simulations. Migrants from one species to another are rare and the impact of migration can be only detected on certain genes[92], which would result in such large fluctuations. The largest migration rate estimated from the 20 models was 4 migrations per mutation, but the impact of migrants is small due to the large effective population size (over 2 million). However, all the 20 IMa2 models consistently predict non-zero gene flows between *Papilio glaucus* and *Papilio canadensis*.

S9.3 Identification of positively selected sites in proteins

We used the PAML (version 4.7a)[93] package to analyze each sequence alignment in order to identify proteins with positively selected positions. A model that allows the dN/dS (ω) to vary among sites was implemented. The null hypothesis was neutral evolution (M1a model[94] in PAML), while the alternative model (M2a model[94] in PAML) allows some positions to undergo purifying or positive selection.
The neutral evolution model was simulated with the following parameters:
runmode = 0, seqtype = 1, CodonFreq = 2, clock = 0, model = 0, NSsites = 1, icode = 0, fix_kappa = 0, kappa = 2, fix_omega = 0, omega = 2, fix_alpha = 1, alpha = 0, Malpha = 0, ncatG = 4, getSE = 0, RateAncestor = 0, method = 0
The alternative model (positive selection) was simulated with the following parameters:
runmode = 0, seqtype = 1, CodonFreq = 2, clock = 0, model = 0, NSsites = 2, icode = 0, fix_kappa = 0, kappa = 2, fix_omega = 0, omega = 2, fix_alpha = 1, alpha = 0, Malpha = 0, ncatG = 4, getSE = 0, RateAncestor = 0, method = 0
We used the following commands to start PAML runs:
$ codeml [control file with parameters]

Positively selected sites have to satisfy the following criteria: (1) alternative model estimates an ω bigger than 1; (2) alternative model is significantly (95% confidence level) more likely than the null hypothesis in a likelihood ratio test. This procedure identified 1301 proteins with positively selected positions. Since we had multiple sequences from both species, each positively selected site may have one or several of the following attributes: (I) same within species, and different between species, (II) different between species; (III) different within *Papilio glaucus*; (IV) different within *Papilio canadensis*. In each protein, we counted the number of positively selected positions with each attribute and these statistics helped to assess

whether adaptive evolution of a particular protein happened mostly within the species or was a mechanism for the two species to diverge. We discovered that positively selected sites detected by PAML are dominated by positions that are variable within a species, which suggests that positive selection was mainly a mechanism for individuals from one species to evolve adaptively.

S9.4 Identification and analysis of divergence hotspots

For each alignment of orthologous transcripts, we calculated the percent identity between a pair of specimens both at the DNA level and at the protein level. Only 351 out of 8230 are always more similar for specimen pairs within species than between species. This small number of proteins likely has played an important role in the speciation process, and therefore we term them "divergence hotspots".

To understand their function, we performed GO term enrichment analysis with in-house scripts. In our analysis, the GO terms, including all parental GO terms that are associated with the divergence hotspots, were compared with GO terms (including all parental GO terms) associated with all 8230 transcripts being analyzed. Significantly enriched GO terms were selected with a binomial test (P = probability for this GO term to be associated with any protein in the background, m = number of divergence hotspots associated with this GO term, N = number of divergence hotspots). We also performed a False Discovery Rate test and calculated the Q-value[34] for each GO term. Significantly enriched GO (P-value < 0.01) terms were visualized in REVIGO[95] web server.

For validation, we performed similar GO term enrichment analysis using the DAVID Bioinformatics Resources 6.7 web server[96,97]. The basic idea behind DAVID server's analysis is similar to our binomial test, but the DAVID server uses more sophisticated models to link more GO terms to a protein by internal relationships between GO terms. The results of our analysis agree well with those produced by the DAVID server. In particular, the most significantly enriched GO terms with Q-values smaller than 0.2 were all significantly enriched according to the DAVID web server.

S9.5 In-depth analysis of circadian clock proteins

Circadian clock proteins: CLOCK, CYCLE, TIMELESS and PERIOD stand out among divergence hotspots because: (1) A GO term associated with all of them, "eclosion rhythm", is among the most significantly enriched ones: there are 5 proteins associated with this GO term in the entire data set, and 4 of them are divergence hotspots. (2) They represent all the central components for the circadian clock pathway. (3) 3 out of 4 proteins differ significantly more between species than within species (confidence level 99%). (4) Differences in them are likely related to their adaptation to locations with very different latitudes, daylight rhythm and timing of seasons[98,99]. (5) Difference in diapause (conditional *vs.* obligate diapause), which is equivalent to difference in the timing of eclosion, is a main phenotypical difference between *Papilio glaucus* and *Papilio canadensis* and also results in different numbers of breeds per year in nature.

We submitted each of these proteins to the MESSA[100] web server to perform secondary structure prediction[101,102], disordered region prediction[103-106], domain and 3D structure template detection by multiple methods[107-109]. Sequences from different specimens and consensus-based predictions are shown in supplementary Fig. S5a-d. We mapped the intra- and inter- species mutations and the predicted domains to the sequences and to the structure templates displayed in Pymol[110].

S10 Selection of nuclear DNA barcodes for insect identification

S10.1 Selection of nuclear barcode candidates that can distinguish *Papilio glaucus* and *Papilio canadensis* specimens

We selected nuclear barcode candidates from a pool of 22731 exons that are longer than 150 bp and shared among all *Papilio glaucus* and *Papilio canadensis* specimens. Wingless and EF1a, nuclear markers used previously to classify insects, are included in this data set[111]. In addition, the Internal Transcribed Spacers (ITS1 and ITS2) and the mitochondrial barcode encoding part of Cytochrome c oxidase subunit 1 (COI) were identified in the *Papilio glaucus* genome and *de novo* assembled transcriptomes, using the *Bombyx mori* sequences and a *Papilio glaucus* COI barcode sequence obtained from Barcode Of Life Database (BOLD)[112], respectively.

We performed binomial tests (p = maximal intra-species mutation rate, m = minimal inter-species variation rate, N = length of the sequence) to determine if these DNA sequences can distinguish *Papilio glaucus* and *Papilio canadensis* specimens with high confidence. In this test, the COI barcode has a P-value of 3.1E-6, and therefore exons with e-values lower than 3.1E-6 were selected as nuclear barcode candidates. The P-value criterion in a binomial test tends to select long sequences. However, short sequences with a high density of mutations would be an easy indicator to discriminate species, and thus they are more desired. Therefore, exons that have a higher inter-species variation rate than the COI barcode and a P-value below 0.05 were also considered to be nuclear barcode candidates.

S10.2 Selection and validation of nuclear barcodes using other insect genomes

The selection procedure in S10.1 resulted in 41 candidates, and we tested them further using information from other insect genomes. We translated their DNA sequences and checked if their orthologs could be detected in other Lepidoptera genomes and aligned with TBLASTN. Exons were considered to be unsuitable as barcodes for the analyzed species if they possessed any of the following characteristics: had multiple comparable TBLASTN hits ($|\log_{10}(\text{e-value1}/\text{e-value2})| < 1.5$); could not detect a confident ortholog (sequence identity > 40%, coverage > 80%). Sequences that were unsuitable for more than one Lepidoptera species were removed from the candidate list, reducing the number of candidate exons to 24.

There are many pairs of closely related species, which have whole genome sequences available, e.g. from the *Drosophila* and *Anopheles* genus. Barcode candidates were tested for their ability to distinguish these species pairs. We downloaded the genomes for these species from various databases[31,113] (Table S23), and identified orthologs for each candidate barcode by TBLASTN followed by manual verification. The sequences from each close species pair were aligned with MAFFT. The mitochondrial barcode COI sequences were missing in several genomes, and in those cases, they were retrieved from BOLD system[112]. Exons without a clearly identifiable and alignable ortholog in most species or that showed no difference between any of species pairs, were removed from the candidate list, leaving 11 exons as our final proposed nuclear barcodes.

The main motivation for defining a new set of nuclear barcodes was to differentiate closely related species. However, they would be more useful if they could also indicate the evolutionary distances. We compared evolutionary trees based on the COI barcode and the nuclear barcodes proposed here to trees based on whole genome data. 12 *Drosophila* species (*Drosophila sechellia*, *Drosophila simulans*, *Drosophila melanogaster*, *Drosophila grimshawi*, *Drosophila virilis*, *Drosophila mojavensis*, *Drosophila willistoni*, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila ananassae*, *Drosophila erecta* and *Drosophila yakuba*) in the phylogenetic tree in Flybase (http://flybase.org/static_pages/species/sequenced_species.html) and 12 *Anopheles* species (*Anopheles arabiensis*, *Anopheles melas*, *Anopheles epiroticus*, *Anopheles stephensi*, *Anopheles maculatus*, *Anopheles culicifacies*, *Anopheles funestus*, *Anopheles dirus*, *Anopheles farauti*, *Anopheles sinensis*, *Anopheles atroparvus* and *Anopheles albimanus*) in the phylogentic tree (https://agcc.vectorbase.org/index.php/Main_Page)[114] from AGCC Vectorbase were used.

The multiple sequence alignments were made by MAFFT and the trees were generated by PhyML and visualized in FigTree. To quantify the similarity in topology, the TOPD/FMTS (version 3.3)[115] program was used with commands like:

$ perl topd_v3.3.pl -f [input file with two trees] -out [output]

Some *Anopheles* species were randomly removed to make the reference phylogenetic tree based on whole genome data[114] into a binary tree, as required for TOPD/FMTS (remaining ones listed above). The phylogenetic trees that are built on the basis of single nuclear barcodes generally agree with trees on whole genome data better than those based on the COI barcode. In particular, evolutionary trees built on some of the nuclear barcodes proposed here can completely reproduce the topology of trees based on whole genome data (Table 2).

S11 Studies of *Papilio appalachiensis*, the hybrid species

S11.1 Identification and alignment of orthologs

We used the same procedure as described in S9.1, except that the number of shared transcripts was reduced to 7411 due to the inclusion of two more *Papilio appalachiensis* specimens. These 7411 transcripts were used for the following studies.

S11.2 Assignment of *Papilio appalachiensis* transcripts to parental species

Statistical tests were used to assign *Papilio appalachiensis* proteins to the parental species. We assign a transcript to one species (species A) if it is significantly different (p-value1 < 0.05) from all orthologous transcripts from species A (p = maximum intra-species variation rate within *Papilio glaucus* and *Papilio canadensis* specimens, m = minimum mutation number compared to any orthologs from species A, N = length of the transcript) and statistically the same (p-value2 > 0.05) as the other species (species B) in another binomial test (p = maximum intra-species variation rate for *Papilio glaucus* and *Papilio canadensis* specimens, m = maximum mutation number compared to any orthologs from species B, N = length of the transcript). A more stringent test with Bonferroni correction requires p-value1 to be smaller than 0.05/7411 = 6.7×10^{-6}.

We used the number of statistically supported *Pgl*-originated transcripts and *Pca*-originated transcripts to estimate the fraction of *Papilio appalachiensis* genes that are inherited from *Papilio canadensis.* In specimen A, we identified 89 *Pgl*-originated and 232 *Pca*-originated transcripts, indicating that 72% of its genes came from *Papilio canadensis*. Similarly, data for specimen B indicates that it inherited 71% of its genes from *Papilio canadensis*.

S11.3 Interpretation of the hybridization event in a whole genome context

The fact that *Papilio appalachiensis* is much more similar to *Papilio canadensis* than to *Papilio glaucus* in terms of overall sequence identity and number of proteins with confidently assigned origins raise the following question: could the presence of statistically supported *Papilio glaucus*-originated proteins in *Papilio appalachiensis* be simply due to insufficient sampling and random fluctuation, and instead *Papilio appalachiensis* is just a close relative of *Papilio canadensis*? If this is true, one would expect these "*Papilio glaucus*-originated proteins" to be randomly distributed in the genome. In contrast, if *Papilio appalachensis* is an outcome of recent hybridization of *Papilio glaucus* and *Papilio canadensis*, one would expect the statistically supported "*Papilio glaucus*-originated proteins" and other *Papilio glaucus*-like proteins (proteins that show higher similarity to *Papilio glaucus* but do not pass binomial test described above) to be close to each other in the genome due to linkage between genes.

We performed statistical tests to check if the *Papilio glaucus*-like proteins tend to cluster in the genome by comparing their real distribution in the genome to that of random samples. In the first test, we counted the number of *Papilio glaucus*-originated genes (P-value < 0.05) that have other *Papilio glaucus*-originated genes (P-value < 0.05) in the same scaffold. Out of 70 such genes, 32 satisfy this criterion. In 10 million random samples of 70 genes from the pool of 7411 genes under study, none have 32 or more genes (the largest is 26 genes) satisfying

this criterion. This indicates the clustering of *Papilio glaucus*-originated genes is extremely significant with a P-value less than 10^{-7}.

In a second test, we calculated the percent of *Papilio glaucus*-like genes (number of *Papilio glaucus*-like genes/(number of *Papilio glaucus*-like genes + number of *Papilio canadensis*-like genes)) in the neighborhood of the 70 statistically supported *Papilio glaucus*-originated genes or random samples of 70 genes. The definition of neighborhood ranged from immediately adjacent genes to genes that are separated by no more than 9 other genes. Genes that share a higher sequence identity to species A than species B by a certain percent identity cutoff (0.2%, 0.4% and 0.6%) were defined as species A-like genes.

Under all these cutoffs, the percent of *Papilio glaucus*-like genes around the statistically supported *Papilio glaucus*-originated genes is always significantly higher than that around random samples: confidence levels are 98.1% for cutoff 0.6% and 99.99% for cutoffs 0.2% and 0.4%, respectively. Statistical significance tends to be lower for a higher identity cutoff: a higher cutoff results in fewer assignable (*Papilio glaucus*-like or *Papilio canadensis*-like) proteins in the neighborhood, and thus among these proteins, the percentage of *Papilio glaucus*-like or *Papilio canadensis*-like genes could reach 100% in random samples due to random chance. As expected, if we include more proteins by increasing the neighborhood range, the statistical significance for a high identity cutoff (0.6%, as shown in supplementary Fig. S7a) also reaches a higher confidence level of 99.9%.

We performed these two statistical tests for *Papilio canadensis*-originated genes and they clustered in the genome even more significantly (P-value < 10^{-7} for the first test and P-value < 0.0001 for the second test under all cutoffs). The higher significance is expected, because *Papilio appalachiensis* inherited more genes from *Papilio canadensis*. The analysis shows that both *Papilio glaucus*-like genes and *Papilio canadensis*-like genes tend to cluster in the genome, which is in consistent with a model of hybridization followed by a limited number of gene recombination events.

Supplementary References

1 Van Nieuwerburgh, F. *et al.* Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic acids research* **40**, e24 (2012).

2 Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* **99**, 45-56 (1999).

3 http://hannonlab.cshl.edu/fastx_toolkit/.

4 Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).

5 Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and correction of sequencing errors. *Genome biology* **11**, R116 (2010).

6 Kim, E. B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479**, 223-227, doi:10.1038/nature10533 (2011).

7 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18, doi:10.1186/2047-217X-1-18 (2012).

8 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1513-1518, doi:10.1073/pnas.1017351108 (2011).

9 Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research* **24**, 1384-1395 (2014).

10 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

11 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

12 Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818-2824, doi:10.1093/bioinformatics/btn548 (2008).

13 Heliconius Genome, C. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94-98 (2012).

14 Zhang, W., Kunte, K. & Kronforst, M. R. Genome-wide characterization of adaptation and speciation in tiger swallowtail butterflies using de novo transcriptome assemblies. *Genome biology and evolution* **5**, 1233-1245 (2013).

15 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

16 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494-1512 (2013).

17 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36 (2013).

18 Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325-2329 (2011).

19 Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666 (2003).
20 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).
21 Duan, J. *et al.* SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology. *Nucleic acids research* **38**, D453-456 (2010).
22 Xia, Q. *et al.* A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science* **306**, 1937-1940 (2004).
23 International Silkworm Genome, C. The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect biochemistry and molecular biology* **38**, 1036-1045 (2008).
24 You, M. *et al.* A heterozygous moth genome provides insights into herbivory and detoxification. *Nature genetics* **45**, 220-225 (2013).
25 Tang, W. *et al.* DBM-DB: the diamondback moth genome database. *Database : the journal of biological databases and curation* **2014**, bat087 (2014).
26 Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**, 1171-1185 (2011).
27 Zhan, S. & Reppert, S. M. MonarchBase: the monarch butterfly genome database. *Nucleic acids research* **41**, D758-763 (2013).
28 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).
29 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
30 Ashburner, M. & Drysdale, R. FlyBase--the Drosophila genetic database. *Development* **120**, 2077-2079 (1994).
31 St Pierre, S. E., Ponting, L., Stefancsik, R., McQuilton, P. & FlyBase, C. FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic acids research* **42**, D780-788 (2014).
32 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
33 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).
34 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445 (2003).
35 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).
36 Smit, A. F. A. & Hubley, R. (http://www.repeatmasker.org) RepeatModeler Open-1.0. (2008-2010).
37 Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**, 1269-1276, doi:10.1101/gr.88502 (2002).

38 Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358, doi:10.1093/bioinformatics/bti1018 (2005).
39 Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Computers & chemistry* **20**, 119-121 (1996).
40 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467 (2005).
41 Smit, A. F. A., Hubley, R. & Green, P. (http://www.repeatmasker.org) RepeatMasker Open-3.0. (1996-2010).
42 Institute, B. (http://www.broadinstitute.org/annotation/genome/Geomyces_destructans/GeneFinding.html) Gene Finding Methods.
43 Misra, S. *et al.* Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome biology* **3**, RESEARCH0083 (2002).
44 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* **6**, 31 (2005).
45 Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288 (2007).
46 She, R. *et al.* genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**, 2141-2143 (2011).
47 Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC bioinformatics* **7**, 62 (2006).
48 Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 59 (2004).
49 Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research* **33**, W451-454 (2005).
50 Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879 (2004).
51 Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**, 188-196 (2008).
52 Liu, Q., Mackey, A. J., Roos, D. S. & Pereira, F. C. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics* **24**, 597-605, doi:10.1093/bioinformatics/btn004 (2008).
53 UniProt, C. Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* **42**, D191-198 (2014).
54 Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology* **5**, e1000605, doi:10.1371/journal.pcbi.1000605 (2009).
55 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
56 Punta, M. *et al.* The Pfam protein families database. *Nucleic acids research* **40**, D290-301, doi:10.1093/nar/gkr1065 (2012).

57 Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic acids research* **40**, D302-305, doi:10.1093/nar/gkr931 (2012).
58 Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research* **41**, D377-386, doi:10.1093/nar/gks1118 (2013).
59 Pedruzzi, I. *et al.* HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic acids research* **41**, D584-589, doi:10.1093/nar/gks1157 (2013).
60 Sigrist, C. J. *et al.* New and continuing developments at PROSITE. *Nucleic acids research* **41**, D344-347, doi:10.1093/nar/gks1067 (2013).
61 Wu, C. H. *et al.* PIRSF: family classification system at the Protein Information Resource. *Nucleic acids research* **32**, D112-114, doi:10.1093/nar/gkh097 (2004).
62 Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic acids research* **41**, D387-395, doi:10.1093/nar/gks1234 (2013).
63 Attwood, T. K. *et al.* The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database : the journal of biological databases and curation* **2012**, bas019, doi:10.1093/database/bas019 (2012).
64 Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162-1164, doi:10.1126/science.252.5009.1162 (1991).
65 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567-580, doi:10.1006/jmbi.2000.4315 (2001).
66 Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology* **338**, 1027-1036, doi:10.1016/j.jmb.2004.03.016 (2004).
67 Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods* **8**, 785-786, doi:10.1038/nmeth.1701 (2011).
68 de Lima Morais, D. A. *et al.* SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic acids research* **39**, D427-434, doi:10.1093/nar/gkq1130 (2011).
69 Lees, J. *et al.* Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic acids research* **40**, D465-471, doi:10.1093/nar/gkr1181 (2012).
70 Kanehisa, M. Molecular network analysis of diseases and drugs in KEGG. *Methods in molecular biology* **939**, 263-275, doi:10.1007/978-1-62703-107-3_17 (2013).
71 Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472-477, doi:10.1093/nar/gkt1102 (2014).
72 Morgat, A. *et al.* UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic acids research* **40**, D761-769, doi:10.1093/nar/gkr1023 (2012).
73 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* **42**, D459-471, doi:10.1093/nar/gkt1103 (2014).

74 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178-2189 (2003).
75 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).
76 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* **52**, 696-704 (2003).
77 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59**, 307-321 (2010).
78 Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS* **8**, 275-282 (1992).
79 FigTree (http://tree.bio.ed.ac.uk/software/figtree/).
80 Moret, B. M. E., Tang, J., Wang, L. & Warnow, T. Steps toward accurate reconstructions of phylogenies from gene-order data. *Journal of Computer and System Sciences* **65**, 508-525 (2002).
81 Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution* **14**, 685-695 (1997).
82 Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research* **36**, W465-469, doi:10.1093/nar/gkn180 (2008).
83 Zhong, Y. F. & Holland, P. W. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & development* **13**, 567-568 (2011).
84 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
85 BCM-HGSC. (https://www.hgsc.bcm.edu/arthropods/tobacco-hornworm-genome-project) Tobacco Hornworm Genome Project.
86 BCM-HGSC. https://www.hgsc.bcm.edu/.
87 Ferguson, L. *et al.* Ancient expansion of the hox cluster in lepidoptera generated four homeobox genes implicated in extra-embryonic tissue formation. *PLoS genetics* **10**, e1004698, doi:10.1371/journal.pgen.1004698 (2014).
88 Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research* **36**, 2295-2300 (2008).
89 Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 2785-2790 (2007).
90 Bruen, T. PhiPack: PHI test and other tests of recombination. (2005).
91 Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Molecular biology and evolution* **27**, 921-933 (2010).
92 Hey, J. Recent advances in assessing gene flow between diverging populations and species. *Current opinion in genetics & development* **16**, 592-596, doi:10.1016/j.gde.2006.10.005 (2006).

93 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591 (2007).
94 Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution* **22**, 1107-1118, doi:10.1093/molbev/msi097 (2005).
95 Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* **6**, e21800 (2011).
96 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
97 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57 (2009).
98 Pfeuty, B. *et al.* Circadian clocks in changing weather and seasons: lessons from the picoalga Ostreococcus tauri. *BioEssays : news and reviews in molecular, cellular and developmental biology* **34**, 781-790, doi:10.1002/bies.201200012 (2012).
99 Merrow, M., Spoelstra, K. & Roenneberg, T. The circadian cycle: daily rhythms from behaviour to genes. *EMBO reports* **6**, 930-935, doi:10.1038/sj.embor.7400541 (2005).
100 Cong, Q. & Grishin, N. V. MESSA: MEta-Server for protein Sequence Analysis. *BMC biology* **10**, 82 (2012).
101 Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228-235 (2002).
102 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **292**, 195-202, doi:10.1006/jmbi.1999.3091 (1999).
103 Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453-1459 (2003).
104 Cheng, J., Sweredoski, M. & Baldi, P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* **11**, 213-222 (2005).
105 Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology* **337**, 635-645, doi:10.1016/j.jmb.2004.02.002 (2004).
106 Lobanov, M. Y. & Galzitskaya, O. V. The Ising model for prediction of disordered residues from protein sequence alone. *Physical biology* **8**, 035004, doi:10.1088/1478-3975/8/3/035004 (2011).
107 Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic acids research* **32**, W327-331, doi:10.1093/nar/gkh454 (2004).
108 Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960, doi:10.1093/bioinformatics/bti125 (2005).
109 Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research* **39**, D225-229, doi:10.1093/nar/gkq1189 (2011).
110 Schrödinger, L. The PyMOL Molecular Graphics System

111 Wilson, J. J. Assessing the value of DNA barcodes and other priority gene regions for molecular phylogenetics of Lepidoptera. *PloS one* **5**, e10525, doi:10.1371/journal.pone.0010525 (2010).
112 Ratnasingham, S. & Hebert, P. D. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular ecology notes* **7**, 355-364, doi:10.1111/j.1471-8286.2007.01678.x (2007).
113 Lawson, D. *et al.* VectorBase: a data resource for invertebrate vector genomics. *Nucleic acids research* **37**, D583-587, doi:10.1093/nar/gkn857 (2009).
114 Neafsey, D. E. *et al.* The evolution of the Anopheles 16 genomes project. *G3* **3**, 1191-1194 (2013).
115 Puigbo, P., Garcia-Vallve, S. & McInerney, J. O. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics* **23**, 1556-1558 (2007).