SUPPLEMENTAL FIGURE LEGENDS

Figure S1, related to Table 1. Comparisons of GC content, CpG ratios, and DNA methylation potential

Shown are GC content in the genome (**A**) and coding regions (genes, **B**), and CpG ratios in the genome (**C**), and the coding regions (**D**). Values are plotted against the frequency. For the genome, GC content and CpG ratios were calculated using 500-bp sliding windows of genomic sequence. Red, monarch; green, *Bombyx*; blue, *Drosophila;* black, *Tribolium*; purple, *A. mellifera*. Grey squares in **D** show the existence of the corresponding member(s) of DNA methyltransferase (Dnmt) family for each species.

Figure S2, related to Figure 2. Comparison of protein domains

A. Pairwise comparison of the InterPro (IPR)-defined family sizes. Bars indicate the number of significantly differing families between each pair of species; color denotes degree of significance. The significance was determined by the Chi-square test with respect to the predicted number of genes with IPR domains.

B. The ten most prominent expansions (upper) and contractions (lower) of monarch IRP families compared to *Bombyx*, listed in decreasing order of significance.

C. The ten most prominent expansions and contractions of lepidopteran IPR families compared to two non-lepidopteran insect species, *Drosophila* or *Tribolium*. See also **Extended Experimental Procedures** for the definitions of expansion and contraction.

Figure S3, related to Figure 3. Animal CRYPTOCHROME phylogeny Maximum likelihood phylogenetic tree showing the evolution of the type-1 (*Drosophila*-like, red lettering) and type 2 (vertebrate-like, green) CRYs in all the arthropods for which draft genomes are available. The tree was rooted with the *E.coli* DNA photolyase. Boostrap values based on 1000 replicates are represented at the nodes. *A. aegypti* : *Aedes aegypti*; *A. cephalotes*: *Atta cephalotes*; *A. gambiae*: *Anopheles gambiae*; *A. mellifera*: *Apis mellifera*; *A. pisum*: *Acyrthosiphon pisum*; *B. mori*: *Bombyx mori*; *C. floridanus*: *Camponotus floridanus*; *C. quinquefasciatus*: *Culex quinquefasciatus*; *D. melanogaster*: *Drosophila melanogaster*; *D. plexippus*: *Danaus plexippus*; *D. pseudoobscura*: *Drosophila pseudoobscura*; *D. pulex*: *Daphnia pulex*; *E. coli*: *Escherichia coli*; *H. saltator*: *Harpegnathos saltator*; *H. sapiens*: *Homo sapiens*; *L. humile*: *Linepithema*

humile; *M. musculus*: *Mus musculus*; *N. vitripennis*: *Nanosia vitripennis*; *P. barbatus*: *Pogonomyrmex barbatus*; *P. h. humanus*: *Pediculus humanus humanus*; *T. castaneum*: *Tribolium castaneum*.

Figure S4, related to Figure 1. Major α subunit gene of monarch P type Na^+/K^+ ATPase

A. Schematic of the genome structure of the major sodium/potassium pump α subunit gene. Black boxes indicate exons and alternative splicing patterns, which were manually curated using transcriptome sequence. The fraction of each splicing pattern is shown around the corresponding positions. Asterisk indicates the position of monarch-specific changes.

B. Hypothetical secondary structure of the α subunit. The secondary structure is based on the topology prediction method, TMHMM Server v. 2.0 (Krogh et al., 2001). The predicted extracellular, intercellular, and transmembrane domains were plotted. Eight major hydrophobic (transmembrane) regions are shown as red peaks.

C. Monarch-specific mutations within the α subunit of Na^+/K^+ ATPase. Multiple alignment of the entire sequence revealed only two monarch-specific mutations, Q(Glu)182V(Val) and N(Asn)193H(His), which are indicated by asterisks. The previous work (Holzinger et al., 1992) was based on DNA sequencing only and missed Q182V because of the mis-splicing of CAG in the intron (in red in lowercase) to the coding region. The magnified region of the first extracellular domain shows the correct splicing pattern.

EXTENDED EXPERIMENTAL PROCEDUES

Animals

Monarchs used for genomic DNA isolation were female migrants. One female was caught in October, 2008 near Eagle Pass, Texas, USA (latitude 28°71′N, longitude 100°49′W) by Carol Cullar, and two females were caught in October, 2008 near Greenfield, Massachusetts, USA (latitude 42°59′N, longitude 72°60′W) by Fred Gagnon.

Genomic Features

Illumina SIPES reads were aligned to assembly using Bowtie v0.12.7 (Langmead et al., 2009) to obtain the best alignment per read pair with the "-k 1 --best" option. The

alignment output was then processed by samtools v0.1.15 (Li et al., 2009) to detect single nucleotide polymorphisms using the suggested parameter values. We identified repetitive sequences and transposable elements using RepeatMasker v3.2.9 (http://www.repeatmasker.org) against a *de novo* repeat library that was built by RepeatModeler v1.0.4 (http://www.repeatmasker.org), as well as the arthropod set of Repbase v20090604 (Lowe and Eddy, 1997). Non-interspersed repeat sequences were also identified by RepeatMasker with the "-noint" option. We predicted transfer RNAs (tRNA) on the repeat-masked genome using tRNAscan-SE-1.23 (Lowe and Eddy, 1997). Distribution of GC content was analyzed in 500-bp non-overlapped windows. CpG ratio, CpG[O/E], is defined as CpG[O/E] = P[CpG]/(P[C]*P[G]), in which P[CpG] is the frequency of CpG dinucleotides, P[C] the frequency of C nucleotides, and P[G] the frequency of G nucleotides.

Transcriptome Analysis

To construct the cDNA library for transcriptome analysis, monarchs from all stages of development were used to ensure a good representation of transcripts: 50 one- to two-days old eggs, one second instar larva raised on milkweed plants, one fifth instar larva raised on diet, one five-day old pupa, and male and female adults from both summer (reproductive) and migrant (non-reproductive) butterflies. To avoid plant contaminants, larvae were dissected in 0.5X RNAlater (Ambion) and their guts were emptied. Heads without antennae, legs, thoraces and abdomens from one male and one female of each state (summer or migrant) were used. Antennae were from two males and two females of each state. Male and female migrant butterflies from which heads, legs and thoraces were used were caught in October, 2008 near Eagle Pass, Texas, USA by Carol Cullar, and those from which antennae and abdomens have been used were caught in October, 2008 near Greenfield, Massachusetts, USA by Fred Gagnon. Summer butterflies were either obtained from Edith Smith (Shady Oak Butterfly Farm, Florida, USA) for all tissues except the antennae, which were obtained from butterflies provided by Orley Taylor (Kansas University, USA). All butterflies were housed in the laboratory in glassine envelopes in incubators with controlled temperature (25°C), humidity (70%), and daily lighting conditions (12h light: 12h dark). Each was fed 25% honey every other day for a week or two prior to collections. To confirm the reproductive status of the

butterflies, female abdomens were dissected. Abdomens from reproductively active summer females contained mature oocytes, while those from migrants did not.

Total RNA was extracted from each developmental stage and for each tissue described above using RNeasy extraction kits (Qiagen; RNeasy Mini kit for eggs and antennae; RNeasy Midi kit for heads, legs and second instar larva; RNeasy Maxi kit for thoraces, abdomens, fifth instar larva and pupa). For heads, thoraces, abdomens and larvae, an additional acidic phenol extraction step was added before binding to the column. Equal amounts of RNA from all preparations were pooled and the sample was stored at -80C until further use. PolyA+ RNA extraction, reverse transcription and cDNA library construction were carried out by The National Center for Genome Resources (Santa Fe, New Mexico, USA).

Gene Models

Approximately 5.4 Gb RNA-seq sequence was employed to generate the transcriptome-based gene models using TopHat v1.2.0 (Trapnell et al., 2009) and Cufflinks v0.9.3 (Trapnell et al., 2010). The invertebrate set of NCBI RefSeq proteins was used for homology search by TBLASTN. The high-scoring pairs (HSP) with $E<10^{-5}$ were then processed by genblastA v1.0.4 (She et al., 2009) and gene structures determined by GeneWise v2.2.0 (Birney et al., 2004). Another five homology-based gene sets were developed independently using EXONERATE v2.2.0 (Slater and Birney, 2005) with gene sets of *Bombyx*, *Drosophila*, *A. gambiae*, *Tribolium*, and *A. mellifera* (Table S1G). Our *ab initio* gene sets were generated from five different predictors: AUGUSTUS v2.5 (Stanke et al., 2006), GeneMark v3.9d (Lomsadze et al., 2005), Genscan (Burge and Karlin, 1997), GlimmerHMM v3.0.1 (Majoros et al., 2004), and SNAP v2006-07-28 (Korf, 2004) (Table S1G). To train the predictors, we also manually curated 282 gene models based on unique monarch ESTs (Zhu et al., 2008). All above individual gene models were integrated to a consensus gene set using GLEAN (Elsik et al., 2007) and Maker v2.08 (Cantarel et al., 2008), respectively. We evaluated sensitivity for each gene models using 20 cloned monarch genes and 784 manually annotated monarch genes based on *Bombyx* homology. Because GLEAN was superior to all the other gene sets, our official gene set (OGS1.0) was based on the non-redundant GLEAN models, with

additional removal of genes that were flagged as repeat elements or were not supported by either homology or the transcriptome.

Orthology and Evolution

All used protein sets of other species are listed in Table S1I. First, we removed very short proteins (<30aa) and filtered out redundant splice variants to keep the longest isoform for each protein set. Next, all-against-all protein comparisons were performed using BLASTP with $E<10^{-5}$. We used orthomclSoftware-v2.0.2 to process HSPs and MCL v10-201 (Li et al., 2003) to define the final orthologs, inparalogs, and co-orthologs, following the suggested parameter values. Multiple alignments of protein sequences for each orthology group were performed using Muscle v 3.8.31 (Edgar, 2004) and the conserved blocks of these alignments were extracted using Gblocks v 0.91b (Talavera and Castresana, 2007). Conserved blocks of 1,642 proteins that have single copies in all species were concatenated to 14 super genes with 377,961 amino acids, which were used to quantify the phylogeny of the 12 insect species. The species tree was calculated using PhyML v2.4.4 (Guindon et al., 2010) with the JTT model. The values of statistical support were obtained from 1,000 replicates of bootstrap analyses. Muscle alignments were also processed by pal2nal v13 (Suyama et al., 2006), the resulting codon alignments were subjected to the calculation of synonymous (dS) and non-synonymous (dN) substitution rates with F3X4 codon frequency, using codeml from the PAML package v Jan-09-2011 (Yang, 2007).

Synteny

Bombyx genomic scaffolds were first concatenated with 500-bp Ns to 28 chromosome sequence according to the information shown in SilkDB 2.0 (Wang et al., 2005). Monarch genes were anchored based on the position of the best BLASTP hit found in the *Bombyx* gene set. For mapping long monarch scaffolds (>10 kb), more than half of the genes within a scaffold that show the consensus position is required to determine the corresponding position on *Bombyx* chromosomes. Pairwise whole genome alignment between the monarch and *Bombyx* was performed using LASTZ v 1.02 with HSP chaining (http://www.bx.psu.edu/~rsharris/lastz/). Because of the 'draft' status of the monarch genome, we only focused on micro-synteny, not chromosome-scale rearrangements.

Quantification of Gene Expression

Based on the transcriptome data, we estimated the general expression value for most predicted genes, except for neuropeptide-related genes, which were of short length that was beyond the library size, and antennal chemoreceptors, because of their general low expression and limited expression in specific cell types. Each predicted coding sequence was extended with 500-bp upstream and downstream regions. Paired-end transcriptome reads were mapped to the extended gene set using Bowtie with up to one alignment report per pair. Sequence coverage was defined as D=N*300/L, in which N is the number of mapped pairs of reads, and L is the length of the gene (we estimated the insert size of the RNAseq library as 300 bp). We also mapped the previously identified ESTs (Zhu et al. 2008) to the extended gene models using BLASTN (both $E<10^{-10}$ and identity > 92% are required). Expression levels for summer and migratory states were calculated based on the raw microarray data (GSE14041 of GEOdatabase). The independent two-sample t-test was used to compare expression values between summer and migratory groups in males and females, respectively.

Annotation of Coding Genes

For automatic annotation, we searched the homology by querying the *Bombyx*, *Drosophila*, and NCBI RefSeq invertebrate protein sets, as well as Gene Ontology and KEGG databases. A local run of InterProScan (IPR) search (Hunter et al., 2009) with all implemented methods was also carried out to identify the conserved domains for gene sets of the monarch, *Bombyx*, *Drosophila*, and *Tribolium*. All above databases were updated to April 2011 for annotation. Species-specific expansion/contraction was determined with the significance of pairwise comparison of the IPR-defined family sizes, which was estimated by the Chi-square test with respect to the predicted number of genes with IPR domains. Several IPR families that are usually found in transposons or are problematic for automatic prediction were omitted, including reverse transcriptase, integrase, zinc finger proteins, and olfactory receptors. Species-specific families, which were missed in all other three species, were also not included in the list. For lepidopteran-specific expansion/contraction, species-specific changed families were first excluded and then families were ordered based on the size difference between the sum of genes in the two lepidopteran and the two non-lepidopteran species.

More than 1,000 genes of biological interest were manually annotated, using *Drosophila*, human, and some well-characterized *Bombyx* orthologs available on NCBI GenBank as queries in most cases. Part of the functional information of *Drosophila* homologs was referred to The Interactive Fly (http://www.sdbonline.org/fly/aimain/1aahome.htm) and GenAge (http://genomics.senescence.info/genes/models.html). Genes with incomplete structures or inappropriate concatenation were identified based on multiple alignments by ClustalX 2.1 (Larkin et al., 2007). If the target homology was not identified in the gene set, additional searches in the genome assembly (by TBLASTN) or raw reads (by Bowtie) was carried out to confirm gene loss. Actually, we have not found, to date, any target gene that only exists in the genome and is not represented in the geneset, which confirms the completeness of our gene models.

Circadian Genes

Drosophila and human sets of clock genes were both utilized to BLASTP search the monarch gene set and other arthropod gene sets. In addition to reciprocal blast, an initial round of phygenetic analysis was performed for cryptochrome (CRY) families to remove the members in (6-4)-photolyase and Cry-DASH clades. This method was also used to differentiate timeless and timeout orthologues. Phylogenetic analysis was performed using PhyML.

We identified the monarch pigment dispersing hormone gene (PDH) based on PF06324 domain (PDH domain in Pfam), as this gene is very short and highly divergent in the N-terminal part of the sequence. Because our current transcriptome did not capture the transcript(s) of PDH, we performed additional polymerase chain reaction (PCR) amplication of cDNA to verify its expression in brain. Total RNA was extracted from a male butterfly brain using RNeasy Mini extraction kit (Qiagen) and cDNA was synthesized using SuperScript II reverse transcriptase (Invitrogen). The primers were designed to span a 1.4 Kb intron and the full-coding region of the peptide, as follow: pdhF, 5'-GCTCTCCCAGCTACGAACTCTA-3'; pdhR, 5'-GATATTCCCGCCATAGACTTG-3'. PCR conditions were as follow: after 5 min at 94°C, five cycles of 30 sec at 94°C, 30 sec at 49°C, 45 sec at 72°C, then 35 cycles of 30 sec at 94°C, 30 sec at 52°C, 45 sec at 72°C, then 5 min of final elongation step at 72°C.

Chemosensory Receptors

Because chemosensory receptor genes are difficult to identify from automated predictions, we identified this class of genes in the genome assembly using TBLASTN searches with *Bombyx*, *Drosophila*, and the moth *Spodoptera littoralis* (only for ionotropic receptors) homologs as queries, followed by iteration. For the genomic loci with significant hits ($E<10^{-3}$), we compared all independent gene sets or re-annotated the exons using GeneWise. Multiple alignments of selected protein sequences were performed using ClustalX. The well-aligned regions were analyzed for phylogenetic analysis using protdist software from the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) with 1,000 replicates of bootstrap analysis.

miRNAs

Migrant butterflies were caught in October, 2010 near Eagle Pass, Texas, USA by Carol Cullar. Total RNA was extracted from 10 summer butterflies and from 10 migrants with Trizol (Invitrogen) and equally pooled from each individual of the two sets for two independent miRNA sequencing lanes (summer and migrant). miRNAs separation, library construction, and Illumina sequencing were conducted by Eureka Genomics. Processed small RNA reads were aligned against the monarch genome by Bowtie, allowing one mismatch. Secondary structures were predicted using RNAfold v1.8.4 (Hofacker, 2003). miRNAs were primarily analyzed by miRDeep pipeline (Friedlander et al., 2008) and manually sorted to remove redundancy. Conserved miRNAs were named according to the unified nomenclature system of miRBase release16 (Kozomara and Griffiths-Jones, 2011). Another two rounds of prediction were conducted using miRTRAP v1.0 (Hendrix et al., 2010) and mireap v0.2 (http://sourceforge.net/projects/mireap/) pipelines. Novel miRNAs that were predicted by all three methods were considered as monarch specific. Remaining mapped reads were aligned to monarch gene models and Rfam r10.0 (Gardner et al., 2009) to identify degraded mRNAs and other non-coding RNAs, respectively. The miRNA expression value for each of the two profiles (summer versus migrant) was normalized to the total number of valid RNA sequence reads per profile.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F., *et al.* (2010). Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. Science *330*, 86-88.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res *14*, 988-995.

Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., *et al.* (2010). Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. Science *329*, 1068-1071.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res *18*, 188-196.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., *et al.* (2011). The ecoresponsive genome of Daphnia pulex. Science *331*, 555-561.

Consortium, International human genome sequencing consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Consortium, The honeybee genome sequencing consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. Nature *443*, 931-949.

Consortium, The international aphid genomics consortium (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol *8*, e1000313.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., *et al.* (2009). Rfam: updates to the RNA families database. Nucleic Acids Res *37*, D136-140.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol *59*, 307-321.

Hendrix, D., Levine, M., and Shi, W. (2010). miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol *11*, R39.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res *31*, 3429-3431.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., *et al.* (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science *298*, 129-149.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., *et al.* (2009). InterPro: the integrative protein signature database. Nucleic Acids Res *37*, D211-215.

Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., *et al.* (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A *107*, 12168-12173.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.
Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res *39*, D152-157.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res *33*, 6494-6506.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955-964.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics *20*, 2878-2879.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., *et al.* (2007). Genome sequence of Aedes aegypti, a major arbovirus vector. Science *316*, 1718-1723.

Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W., Gibbs, R., Beeman, R.W., Brown, S.J., Bucher, G., *et al.* (2008). The genome of the model beetle and pest Tribolium castaneum. Nature *452*, 949-955.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., *et al.* (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res *15*, 1-18.

She, R., Chu, J.S., Wang, K., Pei, J., and Chen, N. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. Genome Res *19*, 143-149.
Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

Smith, C.D., Zimin, A., Holt, C., Abouheif, E., Benton, R., Cash, E., Croset, V., Currie, C.R., Elhaik, E., Elsik, C.G., *et al.* (2011a). Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). Proc Natl Acad Sci U S A *108*, 5673-5678.

Smith, C.R., Smith, C.D., Robertson, H.M., Helmkampf, M., Zimin, A., Yandell, M., Holt, C., Hu, H., Abouheif, E., Benton, R., *et al.* (2011b). Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc Natl Acad Sci U S A *108*, 5667-5672.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res *34*, W435-439.

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., Bouffard, P., Caldera, E.J., Cash, E., Cavanaugh, A., *et al.* (2011). The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet *7*, e1002007.
Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res *34*, W609-612.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol *56*, 564-577.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., Chen, J., Yu, G., Yuan, H., Hu, Y., Li, R., *et al.* (2005). SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res *33*, D399-402.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B.G., Ingram, K.K., Falquet, L., Nipitwattanaphon, M., Gotzek, D., *et al.* (2011). The genome of the fire ant Solenopsis invicta. Proc Natl Acad Sci U S A *108*, 5679-5684.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586-1591.