Supporting Information

Body Louse Genome Sequencing Consortium and Kirkness et al. 10.1073/pnas.1003379107

SI Text

Genome Sequencing, Assembly, and Annotation. Despite a history of inbreeding, the sequenced genomes displayed a relatively high level of polymorphism, and it was necessary to use assembly parameters that were less stringent than described previously (1–3). Overlaps were computed at up to 12% error using 14-mer seeds, ignoring mers present >500 times in the trimmed fragments. Unitigs were computed using overlaps with a maximum of 10% error after correcting for sequencing errors. The genome size used when computing the A-statistic was set to 80 Mb, which biased the algorithm to labeling borderline-deep unitigs as unique instead of repetitive (1). This assembly has been deposited with the National Center for Biotechnology Information (NCBI; accession no. AAZO00000000).

Two large scaffolds (>100 kb), each resembling fragments of a bacterial genome, were used to seed the retrieval of all fragments of the endosymbiont genome. Component reads and their mates were searched iteratively against the complete dataset, and then, a final tally of 44,192 reads was assembled independently into a single contig that represents the entire endosymbiont chromosome. The sequences of this chromosome and an associated plasmid have been deposited with NCBI (accession nos. CP001085 and CP001086).

The *Pediculus humanus humanus* genome assembly was annotated with gene models derived from the VectorBase and JCVI annotation pipelines (4). The initial automated analyses identified 5,797 (VectorBase) and 11,143 (JCVI) gene models. These were merged to yield 10,773 models that were annotated manually by experienced curators (NCBI accession nos. EEB09810–EEB20584). Where genes from disparate sets were mapped to the same genomic locus, the gene with the greatest homology to another insect protein or the longest encoded protein sequence was chosen. Manual annotation was applied only to remedy obvious errors, such as split or merged gene structures or genes targeted based on putative function. The endosymbiont genome was annotated using the JCVI prokaryotic annotation pipeline (http://www.jcvi.org/cms/research/projects/annotation-service/) with manual annotation using the Manatee tool (http://manatee.sourceforge.net/).

To detect GC composition, we partitioned the genomic sequences into segments by the binary recursive segmentation procedure, D_{JS}, proposed by Bernaola-Galván et al. (5). In this procedure, the chromosomes are recursively segmented by maximizing the difference in GC content between adjacent subsequences. The process of segmentation was terminated when the difference in GC content between two neighboring segments was no longer statistically significant (6).

Superscaffolding. We attempted to extend the automated superscaffolding of the 10 largest superscaffolds or groups by manual methods that used all available additional bioinformatic evidence. We were able to make additional links from both ends of most superscaffolds or groups, primarily by using 4-kb mate pairs as custom short contigs that served as stepping stones into the next available large scaffold or group; additionally, we used 10-kb mate pairs and one gene model (40-kb fosmid mate pairs seem to have been exhausted for this purpose).

Telomeres. We searched the trace-archive reads with 1,000 bases of TTAGG repeats, which are the canonical telomeric repeats for insects (7, 8). The first 250 matches among the 9,897 ~40-kb fosmid end reads were plus/minus, indicating that the sequence represented the ends of telomeres. The internal mate pairs of the 70 top-matching reads were almost all repetitive sequences, including some with TTAGG repeats interrupted by non-LTR retrotransposons of the sequence associated repeat telomeric (SART) family, which are also inserted into the telomeres of *Tribolium castaneum* and *Bombyx mori* (9). These insertions almost always occur between the TTA and GG of a telomeric repeat with the poly-A tail oriented to the telomere.

Subtelomeric Structure. A general schematic that is a composite of the structure derived from comparison with the assembled regions of nine telomeres is listed below. The order of telomere components was unique sequence, louse subtelomeric repeat (LSTR1) repeats, short A-rich repeats, LSTR2 repeats, pseudogenes, LSTR1 repeats, unique sequence, SART/TTAGG repeats. This was best exemplified by the 16-kb region at the 3′ end of supercontig 1103172107644, which is telomere 4 below. In the available assembled telomeres, the 5′ end of the subtelomeric region adjacent to unique flanking DNA consists of 5–16 satellite-like repeats of 141 bp (although many have internal regions of these repeats missing so that the repeat length itself is highly variable) called LSTR1 (representative LSTR1: TTTTTTTTTCTTCGTGTTCGTTCCCTCGGTGCAATTGTGCCTCTGTTGCACTGATCGAATCTCGACGCACGTTCAGTTTTTACCGTACGCTCTCGGTCTCGGTCTAGCTCTCGCGCTCGCTCACGCGCTCGATCCCCGGAC). This is followed by 1–2 kb of short A-rich repeats, such as TCCAAAATCAAAATCGAAATCAAAATCGAAATCGAAATTTAAAA. The next 0.5–1 kb consists of runs of thymines (e.g., TTTGGTTTTTTTTTTTTTTTGGATTGGTTTTTTTTTT). This is followed by 4–10 copies of 123-bp LSTR2 (representative LSTR2: CGCGCCCTCCCCCACCCCCACCCGAAACCGCGAGATCGCGGCTCCCGTCGCGGGGTCCGCGTCCGACTTCGGAGAGTCCGGGACCGCGGTCGAAATCCCGAAAAAAAAAAAAAAATTTTTTTT). The next ~8-kb region consists of a unique but shared sequence on each of 4–7 available telomeres and includes several different short pseudogenic regions with best matches in GenBank to genes from monkeys, plants, sea anemone, and fungi. This is followed by a few more LSTR1 repeats. Unfortunately, the highly repetitive nature of these regions has prevented us from manually assembling the connection from this to the SART/TTAGG repeats that must be telomeric of these assembled subtelomeric regions.

The nine assembled telomeric regions are shown below (there are many other small contigs with matches to these that might represent the remaining unassembled telomeres):

Telomere 1: 2 kb at the 3′ end of 57-kb contig 1103172085190 (AAZO01005576.1) that is the 3′ end of 190-kb supercontig 1103172108237 (a singleton Group104). It contains seven LSTR repeats and the short A-rich repeats.

Telomere 2: 10 kb at the 3′ end of 35-kb contig 1103172096746 (AAZO01004088.1) that is the 3′ end of 772-kb supercontig 1103172107761 (Group 19.06). It is the 3′ end of 2.3-Mbp group 19. It contains thymine runs, eight LSTR2 repeats, and the pseudogene region (LSTR repeats and short A-rich repeats are replaced by yet another repeat between flanking unique DNA and the subtelomere).

Telomere 3: 7 kb in reverse orientation at the 5′ end of 73-kb contig 1103172096872 (AAZO01004393.1) that is the 5′ end of 203-kb supercontig 1103172107841 (a singleton Group101). It contains eight LSTR1 repeats, short A-rich repeats, thymine runs, six LSTR2 repeats, and 3 kb of the pseudogenic region.

Telomere 4: 16 kb at the 3′ end of 12-kb contig 1103172096328 (AAZO01003110.1) and all of 7-kb contig 1103172094794 (AAZO01003111.1), which is the 3′ end of 409-kb supercontig 1103172107644 (Group18.02); it is the 3′ end of the largest manual supergroup that is 9 Mbp, the expected length of a chromosome. It contains 16 LSTR1 repeats, short A-rich repeats, thymine runs, 8 LSTR2 repeats, the pseudogenic region, 3 more LSTR1 repeats, and then, 2 kb shared only with short contigs.

Telomere 5: 2 kb at the 3′ end of 28-kb contig 1103172086120 (AAZO01007175.1) that is the 3′ end of 130-kb supercontig 1103172108311 (a singleton Group114). It contains 10 LSTR1 repeats and a few short A-rich repeats.

Telomere 6: 2 kb at the 3′ end of 10-kb contig 1103172095607 (AAZO01001589.1) that is the 3′ end of 393-kb supercontig 1103172107481 (Group10.09); it is the 3′ end of a 4.2-Mbp manual supergroup. It contains seven LSTR1 repeats and short A-rich repeats.

Telomere 7: 7 kb in reverse orientation of 7-kb contig 1103172096930 (AAZO01004592.1) that is the 5′ end of 153-kb supercontig 1103172107879 (Group71.01); it is the 3′ end of a 4.4-Mbp manual supergroup. It contains eight LSTR1 repeats, short A-rich repeats, thymine runs, four LSTR2 repeats, and 3 kb of the pseudogenic region.

Telomere 8: 13 kb at the 3′ end of 48-kb contig 1103172095993 (AAZO01002377.1) that is the 3′ end of 264-kb supercontig 1103172107555 (singleton Group83). It contains six LSTR1 repeats, short A-rich repeats, thymine runs, nine LSTR2 repeats, the pseudogenic region, and ends with two more LSTR1 repeats.

Telomere 9: 1 kb in reverse orientation at the 5′ end of 10-kb contig 1103172094700 (AAZO01003607.1) that is the 5′ end of 56-kb supercontig 1103172107714; it is not in a group. It contains only seven LSTR1 repeats.

Hawkeye Analysis. The genome was assembled by numerous trace reads. However, some important information about the trace reads is often masked in the final analysis. Thus, the compression–expansion (CE) statistic (10, 11) is one way to bridge the gap between the complexity of all of the trace reads from the genome and the linear consensus sequence that is the result of the assembly process. An evaluation of the CE statistic as a predictive measure of misassemblies can be found in Choi et al. (12). The CE statistic compares the implied distance between mate pairs in the assembly with their expected distance based on the clone library size. The CE statistic is defined as the number of SEMs by which a group of insert lengths differs from the expected library mean, and it was calculated by the AMOS software package version 2.0.0 (http://amos.sourceforge.net). We used tools from the AMOS software package (10) to calculate the CE statistic across the genome.

To perform the hawkeye analysis, we used the following approach. If the average inferred length in a region differed from the expected length, that region was deemed suspect using the CE statistic (11). After considering each base position in the assembly, AMOS reported features, or contiguous regions, where the CE statistic indicated that the average insert length in that region was more than three SEs away from the mean. These features represent expansion or compression, depending on if the CE statistic is positive or negative, respectively. Because features are associated with a specific library, a script in the AMOS package, suspfeat2region, was used to combine all features into nonredundant regions of at least 1,000 bp. The detected features and regions will be made available on VectorBase.

We found 5,987 different features, of which 3,810 were expansion features and 2,177 were compression features. Overall, 360 different scaffolds contained one or more features, and 7,770,651 base positions were affected (about 7% of the assembly). By combining overlapping features, we found 4,688 unique regions with a mean of 1.277 features per region. Most of the regions are fairly short with an average length of 2,739 bp. The largest region was 52,099 bp and was located on scaffold 1103172107574. The mean GC content of suspicious regions was 28%, and the mean repeat content was 27%.

The results can be made available in one or more tracks in a genome browser. The CE statistic can be displayed for each base position in the genome, and the features and/or regions can be displayed as intervals. Researchers looking at specific regions of the genome can use these tracks to get some evidence of possible misassemblies in regions of interest. These regions should be used as one piece of evidence and not as absolute predictions.

Transposable Element (TE). We performed two different analyses to identify TEs. In the first analysis, we compared all of the nucleotide sequences of the nonannotated elements with a database of representative sequences extracted from TEfam (http://tefam.biochem.vt.edu/tefam) and the elements previously identified in the genome of *P. h. humanus* by the TE annotation group. These comparisons were made using blastn. In a second analysis, we translated all of the nucleotide sequences of the nonannotated elements using BioEdit, and all of the putative ORFs were compared with a database of representative sequences extracted from TEfam and the elements previously identified in the genome of *P. h. humanus* by the TE annotation group. These comparisons were made using tblastn. These two sets of results showed that the previous set of nonannotated elements correspond mainly to degenerated copies of the previously identified elements (those in the genome paper). The virtual absence of similarity of a subset of short sequences (those shorter than 1,000 bp) with the database of TEs generated led us to hypothesize that most of the shorted nonannotated sequences correspond to remnants of highly degenerated copies of antique TEs.

Representative amino acid sequences were extracted from Repbase (http://www.girinst.org), TEfam (http://tefam.biochem.vt.edu/tefam), and GenBank (http://www.ncbi.nlm.nih.gov). Putative TEs were identified from the *P. h. humanus* genome using an iterative method specific to each class of TEs as outlined below. The TE count in Table 1 represents the number of all blastn hits to the genome with an e value less than 1E-20; this technique was used to report the copy number for all types of TEs. Summary data for TEs in *P. h. humanus* are listed in Table S1*B*.

Class I/Non-LTR Transposable Elements. Several representative reverse-transcriptase amino acid sequences for each non-LTR clade were used as queries for local tblastx searches against the genome. Perl scripts were used to extract the best hits (nucleotide) according to e value (≤1E-20) and length (≥1,000 bp). Flanks were added to each side of these extracted sequences, and then, they were used as seeds for local blastn searches against the genome. The best hits (e value ≤ 1E-20 and length ≥ 1,000 bp) were extracted from the resulting file and aligned using DNASTAR SeqMan II). Two major contigs (along with several minor ones) were obtained and manually examined. The consensus sequences from these contigs were used as seeds to do a final blastn against the genome to estimate the copy number of each element. In addition, the reconstructed elements were used in tblastx searches against the protein database on NCBI (all nonredundant GenBank coding sequence (CDS) translations + RefSeq Proteins + Protein Data Bank + SwissProt + Protein Information Resource + Protein Research Foundation) to identify and compare them with known functional domains of annotated elements.

Class I/LTR Transposable Elements. Several representative reverse-transcriptase amino acid sequences for each of the Ty3/gypsy, Pao/Bel, and Ty1/Copia families were used as queries for local tblastn searches against the genome. Results within 100 bp of one another were combined, and the resulting sequences of length longer than 500 bp were extracted with flanking regions of 3,500 bp. These sequences were used as seeds for blastn and tblastx searches. Results from these searches were used to perform phylogenetic analysis; RNaseH and integrase domains were added to each element, and then, ClustalW was used to perform profile alignments with the alignments as base (13).

Class I/Miniature Inverted-Repeat Transposable Element (MITE) Transposable Elements. A Perl script was used to identify potential MITEs from the genome. This script identified inverted terminal repeats (ITRs) that were at least 11 bp long, not mismatched, no less than 90 bp, and no more than 650 bp apart. ITRs that appeared more than 10 times in the genome were identified, and sequences, including the corresponding ITR, were extracted from the putative MITE ITR. These sequences were then aligned in DNASTAR SeqMan II.

Class II Transposable Elements. Transposase sequences typical to each family were used to perform local tblastx (or tblastn) searches against the genome. A script combined hits within 50 bp of one another, identified results that were of appropriate length (typically two thirds of the transposase length), and then extracted the DNA sequences from the genome with flanking regions appropriate to the length of each element. These data were used for a blastn search to extract the best hits from the results, and DNASTAR SeqMan II was used to align these sequences.

Tandem Repeats. We estimated the content of tandem repeating sequences in both body louse and fruit fly genomes using Tandem Repeats Finder (version 4.04) software (14) with the following parameters: 2 7 7 80 10 50 2000 and the cutoffs as given in Merkel and Gemmell (15).

G Protein-Coupled Receptors. Putative *P. h. humanus* G protein-coupled receptors (GPCRs) were identified by tblastn searches of the louse genome assembly at VectorBase (http://www.vectorbase.org/index.php). The primary source of query sequences included GPCRs from the mosquitoes *Anopheles gambiae* (16) and *Aedes aegypti* (4) as well as *Drosophila melanogaster* (FlyBase; http://flybase.org/), whereas additional invertebrate and vertebrate GPCR sequences were used when appropriate. Manual annotation was performed using Artemis software (Release 7; The Sanger Institute). Alignments of conceptual GPCR amino acid sequences were conducted with ClustalW or MultAlin software (http://bioinfo.genotoul.fr/multalin/multalin.html). Manual annotations were compared with automated gene models (PhumU1.1 gene build) available at VectorBase and also were used to search the *P. h. humanus* genome iteratively for additional GPCR sequences. GPCRs were tentatively categorized according to class and family based on sequence similarity to invertebrate and mammalian GPCRs and named according to nomenclature guidelines developed for invertebrate vectors as detailed at VectorBase. Short peptides presumably representing partial gene models were identified. They may represent gene predictions in regions where errors occurred during the *P. h. humanus* genome assembly, but it was not possible to produce full-length annotations. The *P. h. humanus* nonsensory and opsin GPCRs described in this publication will be made available as third-party annotations through VectorBase.

Odorant-Binding Proteins and Chemosensory Proteins. The identification of the odorant-binding protein (OBP) and chemosensory protein (CSP) genes was performed as in Vieira et al. (17). Briefly, we searched the predicted proteome using blastp and Hidden Markov Model software package (HMMER), and this was followed by a search of the genomic sequence using tblastn. All known OBPs and CSPs were used as query in both blast searches and the PFAM profiles for OBP (PF01395) and CSP (PF03392) in HMMER searches. All results were manually curated, and the putative gene structure was checked for known OBP/CSP characteristics (signal peptide, typical secondary structure, presence of start and stop codons, etc.).

P450, GST, and EST genes. The peptide sequences of well-characterized representative genes from *D. melanogaster*, *An. gambiae*, *A. mellifera*, and *T. castaneum* were used as queries to search the louse genome database at VectorBase (http://phumanus.vectorbase.org/) by blastp. Groups of a target gene family exhibiting highly significant matches (mostly >40%) were retrieved, and then, using the *P. h. humanus* sequences as queries in turn, the PhumU1.1 peptide database blastp search was repeated until no new target genes were found. After putative target gene sets were identified from the human body louse genome, they were subsequently used as queries for the NCBI blastp search to verify their identity and phylogenetic relationships with other known genes.

Insulin/Target of Rapamycin (TOR) Pathway Genes. To analyze the body louse insulin/TOR pathway genes, the orthologs of the *D. melanogaster* insulin/TOR genes in the *P. h. humanus* genome were identified using a best reciprocal blast approach (18). Each candidate gene was evaluated manually. Gene structure was determined using information from multiple sequence alignment of known insect insulin/TOR pathway genes and, when available, the *Pediculus* predicted transcripts and EST information. For identification of the insulin-like peptide genes, we used the characteristic amino acid pattern (a number of cysteines spaced by a specific number of residues) (19, 20) observed in vertebrates and most invertebrate species.

Interestingly, the body louse has orthologs for all *D. melanogaster* insulin/TOR pathway genes (Dataset S2*D*), and therefore, the body louse genome would encode a complete and functional insulin/TOR pathway. However, the number of genes was lower in the body louse than in *D. melanogaster*. Indeed, in *D. melanogaster*, 14 insulin/TOR pathway genes are single copy, whereas the rest belong to two paralogous groups: seven genes encode the *Drosophila* insulin-like peptides (*dilp1–7*), and another seven genes encode the elongation initiation factor 4E (*eIF-4E*, *eIF4E3–7*, and *4EHP*). In contrast, the *P. h. humanus* genome contains a single insulin-like peptide and three eIF4E-encoding genes. All three eIF4E gene classes described in Joshi et al. (21) were represented in the *P. h. humanus* genome, whereas class III is missing in Diptera.

Nonreduced Gene Families. ***Nuclear receptor superfamily genes.*** Members of the nuclear receptor (NRs) super family share a characteristic modular structure with the DNA-binding and ligand-binding domains being the most widely conserved among different NRs (Dataset S2 *G* and *H*). Most of the NRs act as ligand-activated transcription factors (22), mediating between signaling molecules like hormones and transcription factors that regulate spatial and temporal expression of genes involved in various developmental processes (23–25). Using the amino acid sequences of C4-Zn finger domain and ligand-binding domain in the blast search tool, we have identified 22 putative NRs (of which 20 are orthologous to the NRs in *D. melanogaster*) and 1 NR gene (PHUM8965) with incomplete sequence in the body louse genome (Dataset S2 *G* and *H*). Of 21 NRs in *D. melanogaster*, only 1 gene HR83 (NR2E5, FBgn0037436) was not found in the body louse genome.

Channel and receptor super-family genes. The following *P. h. humanus* neuronal component genes were found to be highly conserved among insects: (*i*) voltage-dependent sodium-channel α-subunits (VDSC), (*ii*) sodium-channel auxiliary subunits, and (*iii*) nicotinic acetylcholine-receptor subunits (nAChR). Using amino acid comparisons, two

VDSC genes orthologous to para and NCP60E (CG9071) sodium channels from *D. melanogaster* were identified in the *P. h. humanus* genome. These findings are identical to those in other known insect genomes, including *An. gambiae*, *A. mellifera*, and *T. castaneum*, in which single orthologs for each VDSC are present. There are five homologs to the *Drosophila* tipE, known as the insect sodium-channel auxiliary subunit gene, in *P. h. humanus*. Each gene of the tipE family was represented by a single orthologous gene and showed a high degree of conservation with other insects. Nine genes homologous to nAChRs in other insects were found in *P. h. humanus*. The putative nAChR genes were categorized into eight groups (a single gene in each group of Dα1, Dα2, Dα3, Dα4, Dβ1, Dβ2, and more distantly related Dβ3 versus two genes in Dα5–7) (26). Other insects, such as *D. melanogaster*, *An. gambiae*, and *A. mellifera*, have 10 nAChR genes, and their distribution is very similar to that of *P. h. humanus* (26). This similarity in the number and composition of nAChR genes suggests that they are highly conserved across insect taxa, even with remarkably different life history and ecology; this reflects their evolutionarily retained function.

Neurohormones and neuropeptides. Apart from insulin, insects use a number of neurohormones and neuropeptides that act through GPCRs to regulate a variety physiological processes. A large number of these neuropeptides have been identified, and in many cases, their receptors are also known from at least one insect species, usually *D. melanogaster* [review by Hauser et al. (27)]. Although most of insect neuropeptide genes are present in the louse genome (Dataset S2 *G* and *H*), genes encoding proctolin, vasopressin, and allatotropin were missing. These peptides are probably genuinely absent from the genome, because the homologs for their receptors have not been recovered. Both vasopressin and allatotropin are also lacking from *D. melanogaster* (28), whereas proctolin and vasopressin are missing from the *B. mori* genome (29). Thus, the louse genome seems to be relatively complete in regards to the neuropeptide genes, except for these proteins.

Genes associated with wing development. The absence of wings in all extant Phthiraptera (true lice) represents a drastic morphological adaptation to their parasitic lifestyle. The origin of this evolutionary adaptation is quite old, because fossil records and phylogenetic analyses suggest that the Phthirapteran lineage (and the winglessness) probably appeared in the early Cretaceous to late Jurassic (140–150 mya) period (30). Hence, true lice can serve as an excellent system to study the molecular evolution of genes that were responsible for ancestral wing development. One possibility is that the actual loss of these genes in lice led to the subsequent loss of wings. Alternatively, winglessness may have evolved through the modification of the expression pattern of wing genes. Decades of studies in developmental biology suggest that the latter scenario is more likely, because many (if not all) developmental genes have pleiotropic functions and their loss would be detrimental. However, the former scenario might also be possible and is suggested by the loss of a Hox gene in crustaceans with truncated abdomens (31). To begin to understand the molecular basis behind the evolution of winglessness in lice, we have surveyed wing genes in the louse genome. Of more than 30 genes known to be important for wing development in *D. melanogaster*, we could not detect any gene loss in this category. Even crossveinless 2 (cv-2), a gene that has rather minor phenotypic effects in *D. melanogaster*, had a highly conserved louse ortholog. This result indicates that these *Pediculus* orthologs have important functions other than wing development. Thus, the evolution of winglessness in lice has been likely achieved through loss of wing-specific gene expression, possibly by modification of wing-specific *cis*-regulatory elements. Detailed expression analysis for these genes in lice may help us to understand the molecular basis of winglessness in Phthiraptera.

1. Myers EW, et al. (2000) A whole-genome assembly of *Drosophila*.. *Science* 287: 2196–2204.
2. Venter JC, et al. (2001) The sequence of the human genome. *Science* 291:1304–1351.
3. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5: e254.
4. Nene V, et al. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316:1718–1723.
5. Bernaola-Galvan P, Roman-Roldan R, Oliver JL (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 53:5181–5189.
6. Cohen N, Dagan T, Stone L, Graur D (2005) GC composition of the human genome: In search of isochores. *Mol Biol Evol* 22:1260–1272.
7. Okazaki S, Tsuchida K, Maekawa H, Ishikawa H, Fujiwara H (1993) Identification of a pentanucleotide telomeric sequence, (TTAGG)n, in the silkworm *Bombyx mori* and in other insects. *Mol Cell Biol* 13:1424–1432.
8. Traut W, et al. (2007) The telomere repeat motif of basal Metazoa. *Chromosome Res* 15:371–382.
9. Fujiwara H, Osanai M, Matsumoto T, Kojima KK (2005) Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, *Bombyx mori*. *Chromosome Res* 13:455–467.
10. Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: Finding the elusive mis-assembly. *Genome Biol*, 10.1186/gb-2008-9-3-r55.
11. Zimin AV, Smith DR, Sutton G, Yorke JA (2008) Assembly reconciliation. *Bioinformatics* 24:42–45.
12. Choi JH, et al. (2008) A machine-learning approach to combined evidence validation of genome assemblies. *Bioinformatics* 24:744–750.
13. Tubio JM, Naveira H, Costas J (2005) Structural and evolutionary analyses of the Ty3/gypsy group of LTR retrotransposons in the genome of *Anopheles gambiae*. *Mol Biol Evol* 22:29–39.
14. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580.
15. Merkel A, Gemmell NJ (2008) Detecting microsatellites in genome data: Variance in definitions and bioinformatic approaches cause systematic bias. *Evol Bioinform Online* 4:1–6.
16. Hill CA, et al. (2002) G protein-coupled receptors in *Anopheles gambiae*. *Science* 298: 176–178.
17. Vieira FG, Sanchez-Gracia A, Rozas J (2007) Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol*, 10.1186/gb-2007-8-11-r235.
18. Alvarez-Ponce D, Aguade M, Rozas J (2009) Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res* 19:234–242.
19. Claeys I, et al. (2002) Insulin-related peptides and their conserved signal transduction pathway. *Peptides* 23:807–816.
20. Smit AB, et al. (1998) Towards understanding the role of insulin in the brain: Lessons from insulin-related signaling systems in the invertebrate brain. *Prog Neurobiol* 54: 35–54.
21. Joshi B, Lee K, Maeder DL, Jagus R (2005) Phylogenetic analysis of eIF4E-family members. *BMC Evol Biol* 5:48.
22. Oro AE, McKeown M, Evans RM (1992) The *Drosophila* retinoid X receptor homolog ultraspiracle functions in both female reproduction and eye morphogenesis. *Development* 115:449–462.
23. Karin M, Yang-Yen HF, Chambard JC, Deng T, Saatcioglu F (1993) Various modes of gene regulation by nuclear receptors for steroid and thyroid hormones. *Eur J Clin Pharmacol* 45(Suppl 1):S9–S15.
24. Luisi BF, Schwabe JW, Freedman LP (1994) The steroid/nuclear receptors: From three-dimensional structure to complex function. *Vitam Horm* 49:1–47.
25. Wahli W, Martinez E (1991) Superfamily of steroid nuclear receptors: Positive and negative regulators of gene expression. *FASEB J* 5:2243–2249.
26. Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB (2007) Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* 445:86–90.
27. Hauser F, et al. (2008) A genome-wide inventory of neurohormone GPCRs in the red flour beetle *Tribolium castaneum*. *Front Neuroendocrinol* 29:142–165.
28. Taghert PH, Veenstra JA (2003) *Drosophila* neuropeptide signaling. *Adv Genet* 49: 1–65.
29. Roller L, et al. (2008) The unique evolution of neuropeptide genes in the silkworm *Bombyx mori*. *Insect Biochem Mol Biol* 38:1147–1157.
30. Grimaldi D, Engel MS (2006) Fossil Liposcelididae and the lice ages (Insecta: *Psocodea*). *Proc Biol Sci* 273:625–633.
31. Geant E, Mouchel-Vielh E, Coutanceau JP, Ozouf-Costaz C, Deutsch JS (2006) Are Cirripedia hopeful monsters? Cytogenetic approach and evidence for a Hox gene cluster in the cirripede crustacean *Sacculina carcini*. *Dev Genes Evol* 216:443–449.

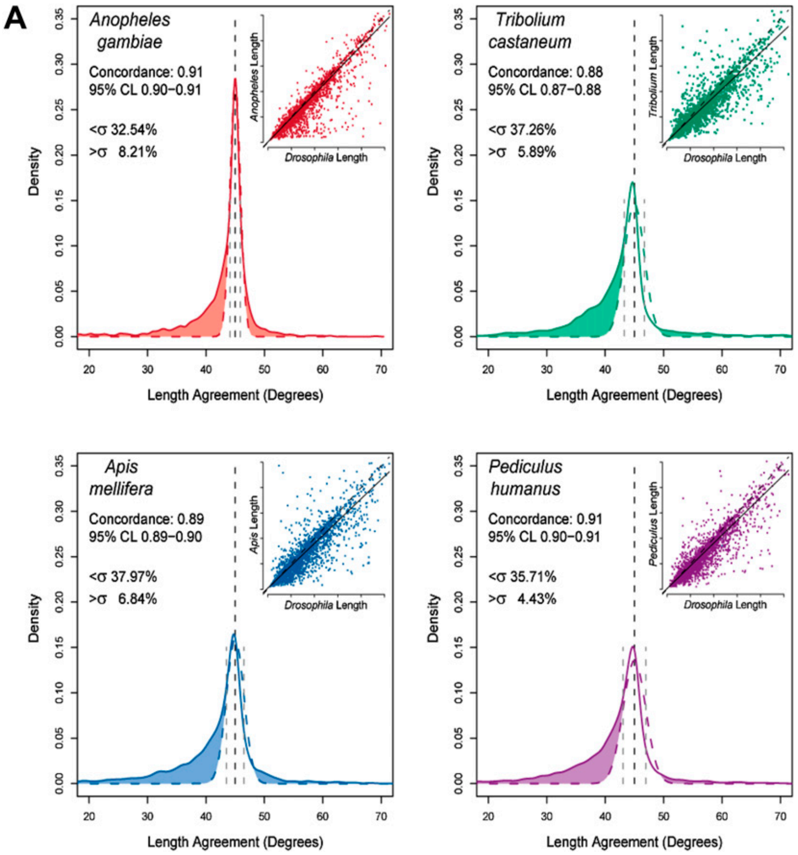


Fig. S1. (Continued)

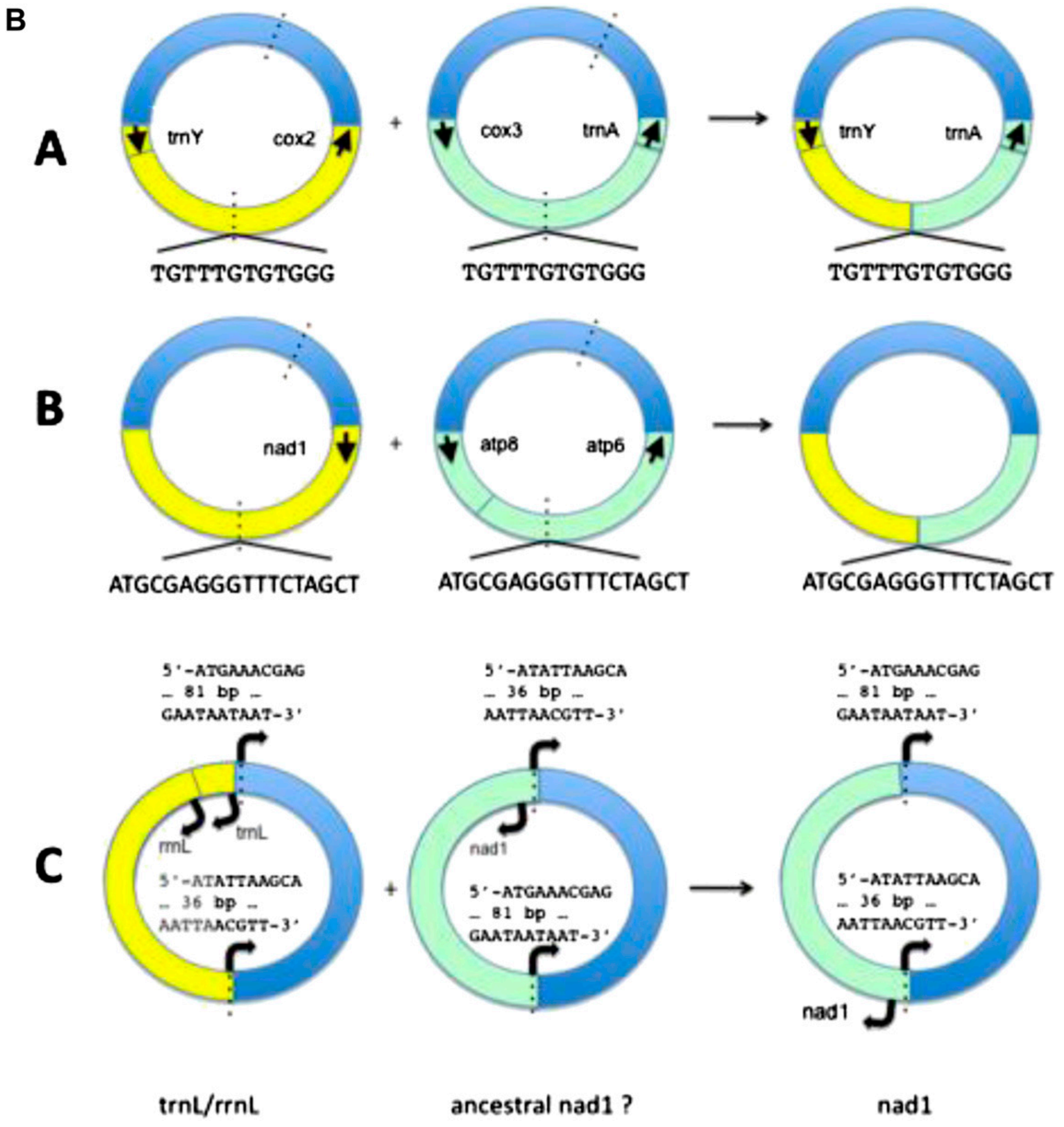


Fig. S1. (*A*) Orthologous protein-length analysis. Orthologous protein-length agreement between *Drosophila melanogaster* proteins with single-copy orthologs in four other insect species: *Anopheles gambiae* (red), *Tribolium castaneum* (green), *Apis mellifera* (blue), and *Pediculus humanus humanus* (purple). The amino acid lengths of 3,753 strict single-copy orthologs (one member in each of the five species) sourced from OrthoDB were compared using the well-annotated *Drosophila* proteins as the baseline. The scatter plots in *Insets* show the *Drosophila* protein length (*x*) against the orthologous protein length (*y*) for each species: axes are from 0 to 2,500 amino acids, the dashed lines show perfect agreement (*x* = *y*; 45°), and the solid lines show a robust linear regression. The concordance of *x* and *y* is given with 95% confidence limits (CL), and perfect concordance (1.0) would require all points to fall on the 45° line. To examine the distributions of evident deviations from perfect agreement, the density of data points falling at each degree below and above 45° is plotted (solid colored curves). These density distributions are compared with normal fittings of the data (dashed colored curves) with means fixed at 45° (dashed black vertical lines). The areas representing the positive differences between the observed data and the normal fitted data below and above 1 SD from the mean of the normal fitted data (σ, dashed gray vertical lines) are filled with the respective colors for each species. The values of these proportions of significantly shorter proteins (<σ) and significantly longer proteins (>σ) are enumerated for quantitative comparisons. *P. h. humanus*, despite being the most distantly related to *Drosophila* of the considered species, exhibits the same level of concordance (0.91) as the much more closely related *A. gambiae* and better concordance than both *T. castaneum* (0.88) and *A. mellifera* (0.89). This is reflected in the proportions of significantly shorter or longer proteins in each of the species comparisons, and this supports the conclusion that, despite the large evolutionary distances from other insects, the *P. h. humanus* protein-coding gene set is remarkably accurate. (*B*) A model of nonhomologous end joining (NHEJ) between mitochondrial minichromosomes that generated chimeric mitochondrial chromosomes in *P. h. humanus*. Coding regions of minichromosomes are in yellow and green, and noncoding regions are in blue. Black arrows in coding regions indicate the orientation of gene transcription. Broken lines indicate sites of double-strand breakages where the two minichromosomes that recombine share homologous sequences. Of 37,144 sequence reads that contained mitochondrial genes, a small number (1.5%) aligned only partially with the 18 abundant minicircular chromosomes. Almost all (98%) of these 529 reads could be assembled into two chromosomes, each a chimeric derivative of two known chromosomes that seem to have recombined by NHEJ through a common microhomologous sequence of 12 bp (*Top*) or 19 bp (*Middle*). The protein-coding genes of the chimeric chromosomes have only fragments of the full-length cox2, cox3, nad1, and atp6 genes. However, the two tRNA genes, trnA and trnY, were the same length as their counterparts in the known minichromosomes and therefore, potentially functional. Interestingly, the genic regions of all mitochondrial chromosomes have a common upstream motif (CAAAYCTCAACTCGTTTCAT), and all except one have the same orientation relative to the conserved noncoding region (23). The exceptional chromosome (encoding nad1) shares a 56-bp segment with rrnL that may have arisen from a similar NHEJ event between the ancestral nad1 and rrnL minichromsomes (*Bottom*).

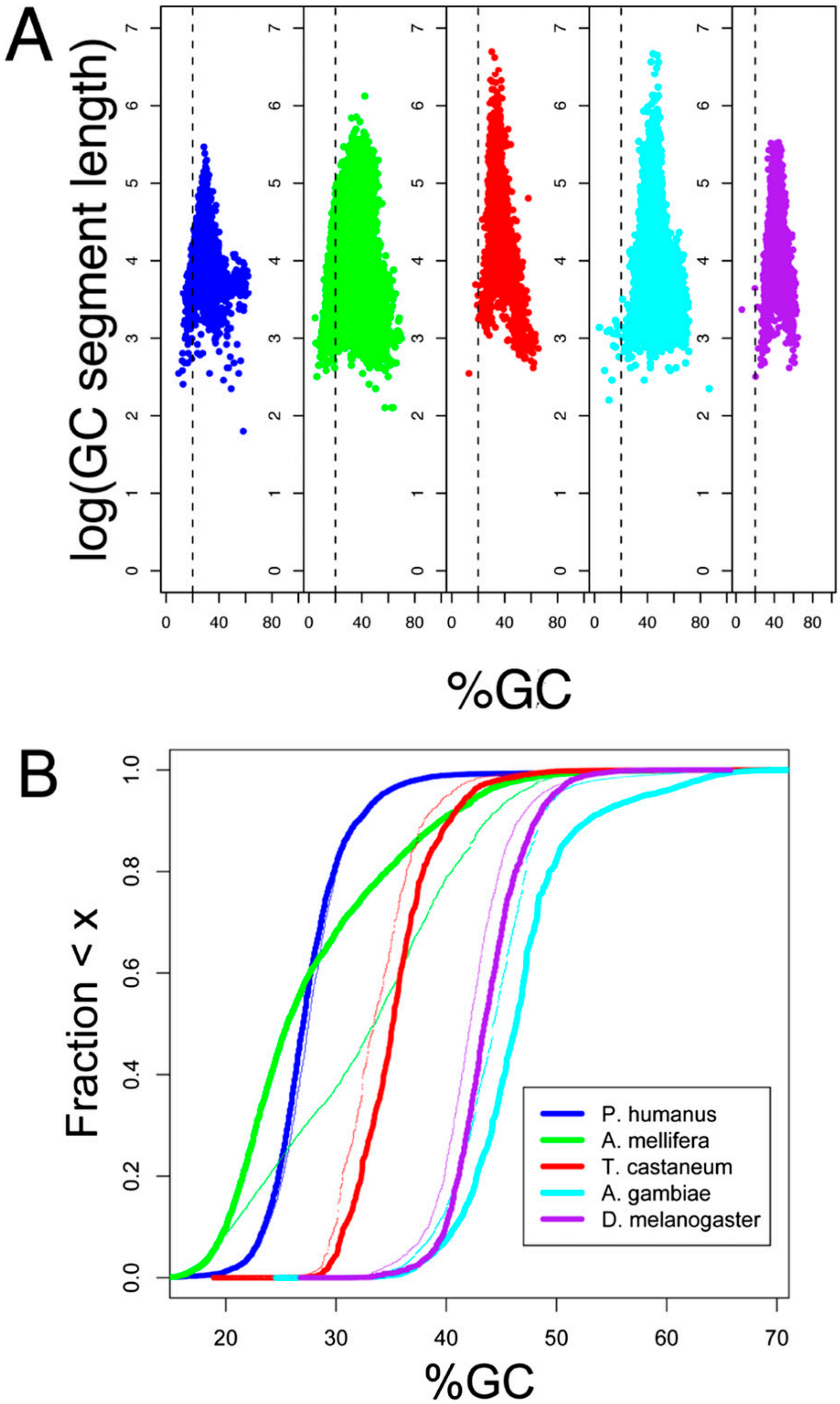


Fig. S2. Comparison of GC-content domains in the insects *Pediculus humanus humanus*, *Apis mellifera*, *Tribolium castaneum*, *Anopheles gambiae*, and *Drosophila melanogaster*. (*A*) GC-content domain lengths versus GC percentage. Hatched line at 20% shown for comparison. (*B*) *P. h. humanus* genes show a very slight tendency to occur in AT-rich regions of the genome. Cumulative distributions show the fraction of genes (thick lines) or the entire genome (thin lines) occurring in GC-content domains (<*x* GC%).

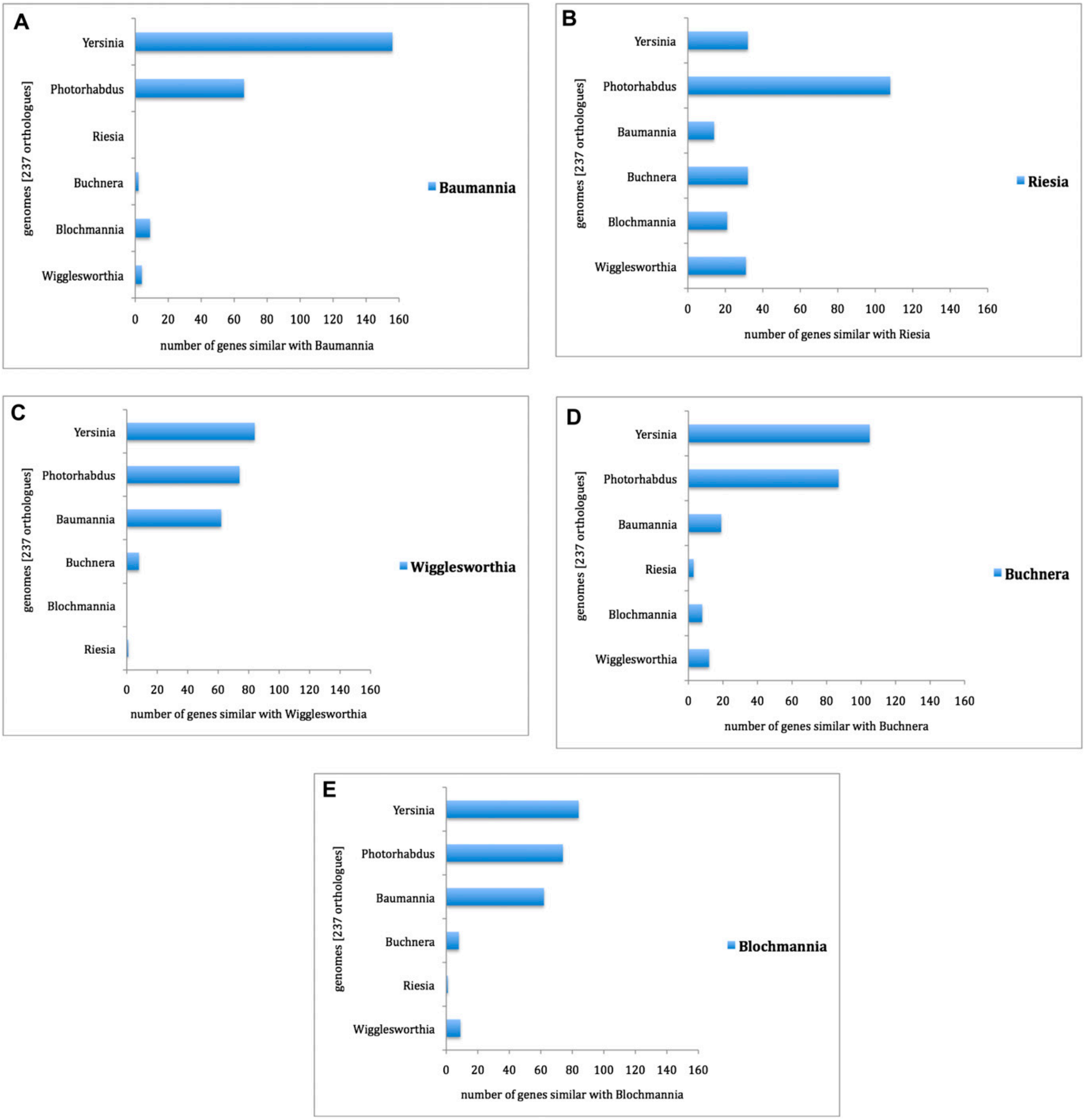


Fig. S3. A genome-wide comparison of *Candidatus* Riesia pediculicola with the primary endosymbionts, *Wigglesworthia glossinida* (tsetse flies), *Blochmannia floridanus* (not shown), *B. pennsylvanicus* (carpenter ants), the automonous *Buchnera aphidicola* (aphids) strains APS and BBp, Sg (not shown), *Baumannia cicadellinicola* (leafhoppers and sharpshooters), and the pathogens, *Photorhabdus luminescens* subsp. laumondii TTO and *Yersinia pestis* str. CO92, revealed a core of 237 genes in all aforementioned bacteria with only 27 genes unique to Riesia (Table S2*B*) and 30 genes present in all bacteria except Riesia (Table S2*A*). In this comparison, Riesia shares the most orthologs with *P. luminescens.*

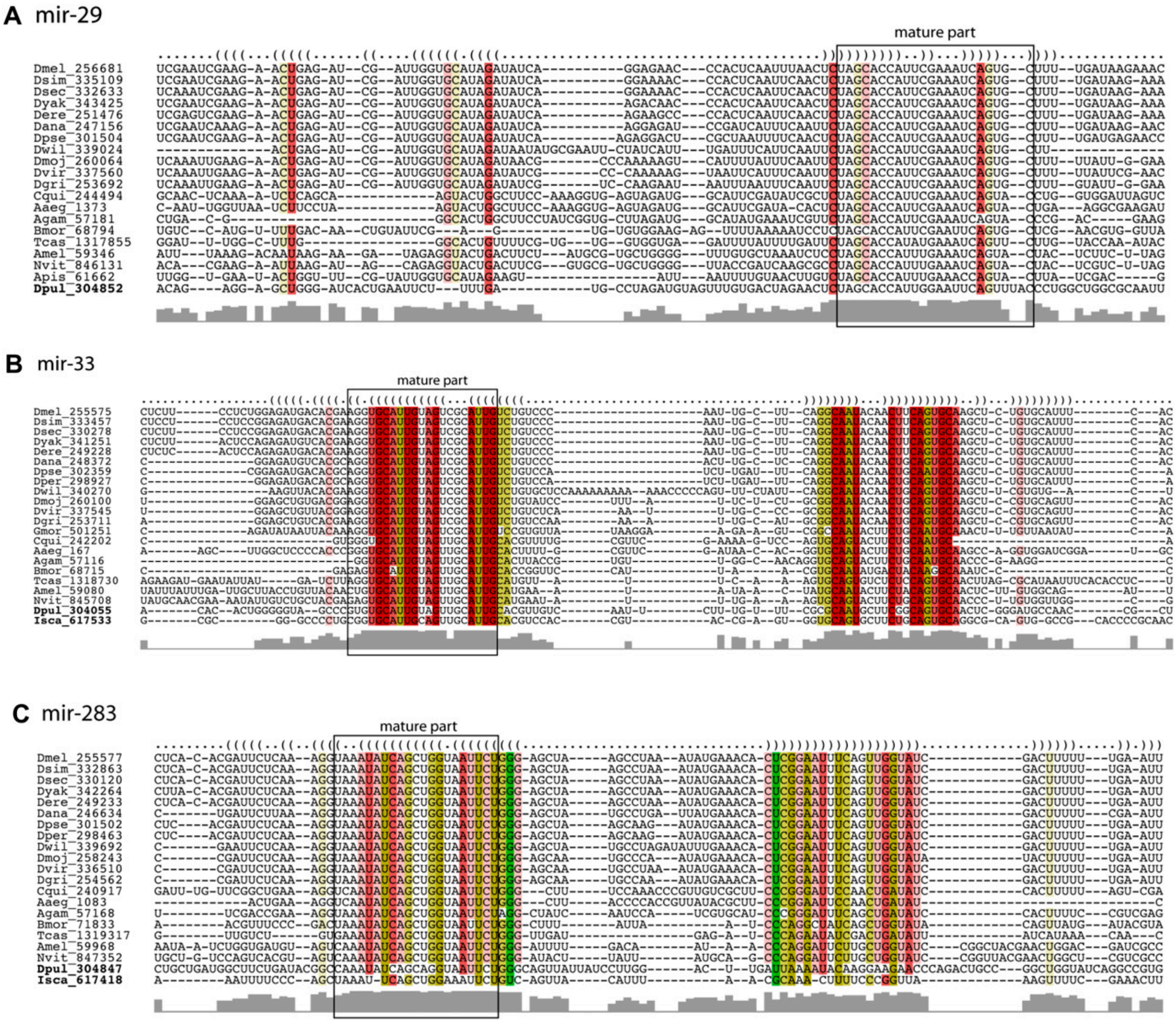
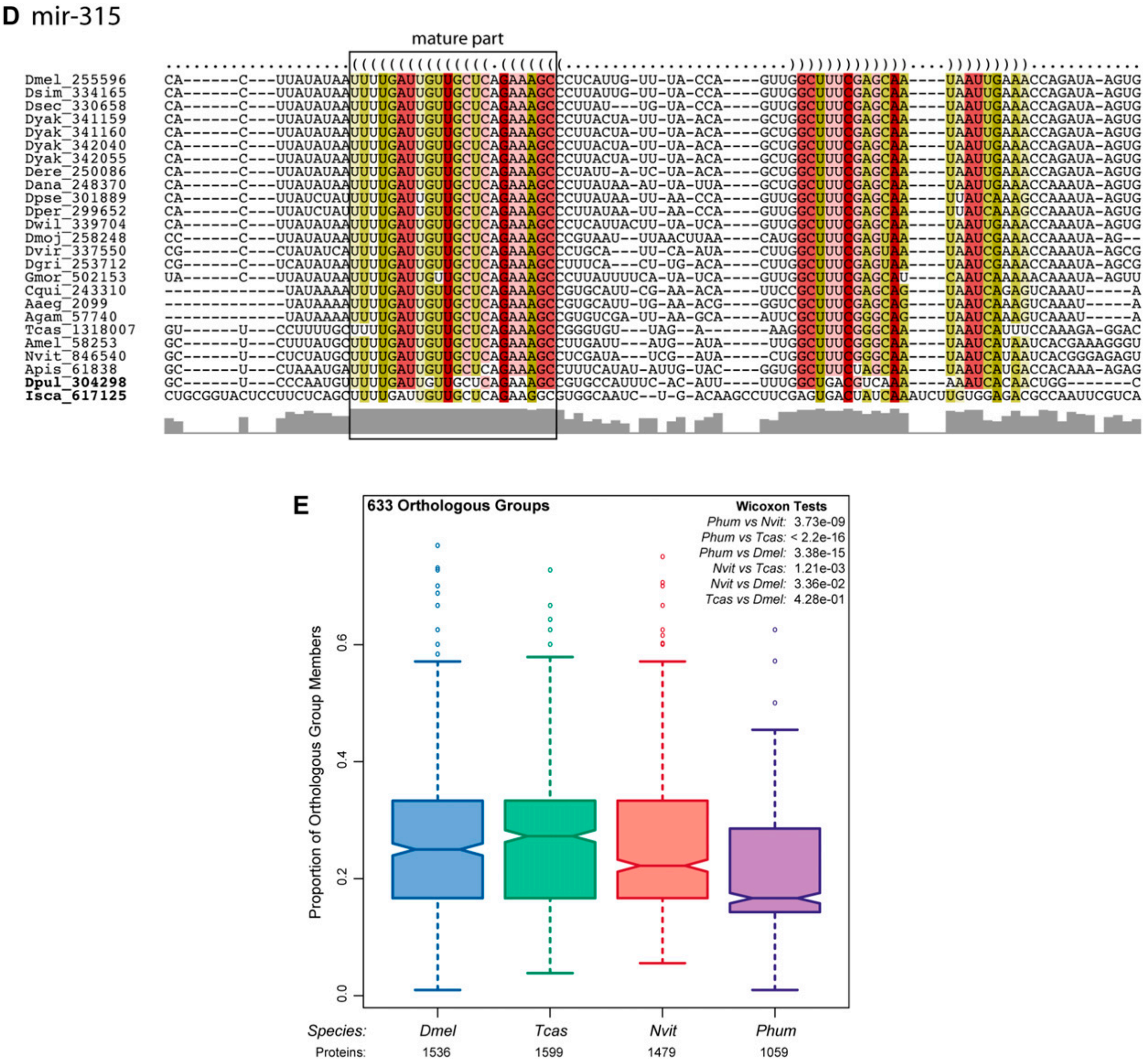


Fig. S4. (Continued)



```
               ....................(((((((((((((((.((((((((........................))))))))))))).....)))))))))............
Dmel_255596    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUCAUUG-UU-UA-CCA----GUUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dsim_334165    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUAUUG-UU-UA-CCA----GUUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dsec_330658    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUAU---UG-UA-CCA----GUUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dyak_341159    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUACUA-UU-UA-ACA----GCUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dyak_341160    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUACUA-UU-UA-ACA----GCUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dyak_342040    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUACUA-UU-UA-ACA----GCUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dyak_342055    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUACUA-UU-UA-ACA----GCUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dere_250086    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUAUU-A-UC-UA-ACA----GCUGGCUUUCGAGCAA----UAAUUGAAACCAGAUA-AGUG
Dana_248370    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUAUAA-AU-UA-UUA----GCUGGCUUUCGAGCAA----UAAUUGAAACCAAAUA-AGUG
Dpse_301889    CA------C---UUAUCUAUUUUUGAUUGUUGCUCAGAAAGCCCUUAUAA-UU-AA-CCA----GUUGGCUUUCGAGCAA----UUAUCAAAGCCAAAUA-AGUG
Dper_299652    CA------C---UUAUCUAUUUUUGAUUGUUGCUCAGAAAGCCCUUAUAA-UU-AA-CCA----GUUGGCUUUCGAGCAA----UUAUCAAAGCCAAAUA-AGUG
Dwil_339704    CA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUCAUUACUU-UA-UCA----GUUGGCUUUCGAGCAA----UAAUUGAAACCAAAUA-AGUG
Dmoj_258248    CC------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCGUAAU--UUAACUUAA----CAUGGCUUUCGAGUAA----UAAUCGAAACCAAAUA-AG--
Dvir_337550    CG------C---CUAUAUCAUUUUGAUUGUUGCUCAGAAAGCCCUGCA---UU-CA-AUA----CUUGGCUUUCGAGUAA----UAAUCGAAACCAAAUA-AGCG
Dgri_253712    CG------C---UCAUAUAAUUUUGAUUGUUGCUCAGAAAGCCUUUCA---CU-UG-ACA----CUUGGCUUUCGAGUAA----UAAUCGAAACCAAAUA-AGCG
Gmor_502153    UA------C---UUAUAUAAUUUUGAUUGUUGCUCAGAAAGCCCUUAUUUUCA-UA-UCA----GUUGGCUUUCGAGCAU----CAAUCAAAAACAAAUA-AGUU
Cqui_243310    -------------UAUAAAAUUUUGAUUGUUGCUCAGAAAGCCGUGCAUU-CG-AA-ACA----UUCCGCUUUCGAGCAG----UAAUCAGAGUCAAAU-----A
Aaeg_2099      -------------UAUAAAAUUUUGAUUGUUGCUCAGAAAGCCGUGCAUU-UG-AA-ACG----GGUCGCUUUCGAGCAG----UAAUCAAAGUCAAAU-----A
Agam_57740     -------------UAUAAAAUUUUGAUUGUUGCUCAGAAAGCCGUGUCGA-UU-AA-GCA----AUUCGCUUUCGGGCAG----UAAUCAAAGUCAAAU-----A
Tcas_1318007   GU------U---CCUUUUGCUUUUGAUUGUUGCUCAGAAAGCCGGGUGU---UAG--A-------AAGGCUUUCGGGCAA----UAAUCAUUUCCAAAGA-GGAC
Amel_58253     GC------U---CUUUAUGCUUUUGAUUGUUGCUCAGAAAGCCUUGAUU---AUG--AUA-----UUGGCUUUCGGGCAA----UAAUCAUAAUCACGAAAGGGU
Nvit_846540    GC------U---CUCUAUGCUUUUGAUUGUUGCUCAGAAAGCCUCGAUA---UCG--AUA-----CUGGCUUUCGGGCAA----UAAUCAUAAUCACGGGAGAGU
Apis_61838     GC------U---CUAAAUGAUUUUGAUUGUUGCUCAGAAAGCCUUUCAUAU-AUUG-UAC----GGUGGCUUUCUAGCAA----UAAUCAUGACCACAAA-AGAG
Dpul_304298    GC------U---CCCAAUGUUUUUGAUUGUUGCUCAGAAAGCCGUGCCAUUUC-AC-AUU----UUUGGCUGACGUCAAA----AAAUCACAACUGG-------C
Isca_617125    CUGCGGUACUCCUUCUCAGCUUUUGAUUGUUGCUCAGAAGGCGUGGCAAUC--U-G-ACAAGCCUUCGAGUGACUAUCAAAUCUUGUGGAGACGCCAAUUCGUCA
```

Fig. S4. (Continued)

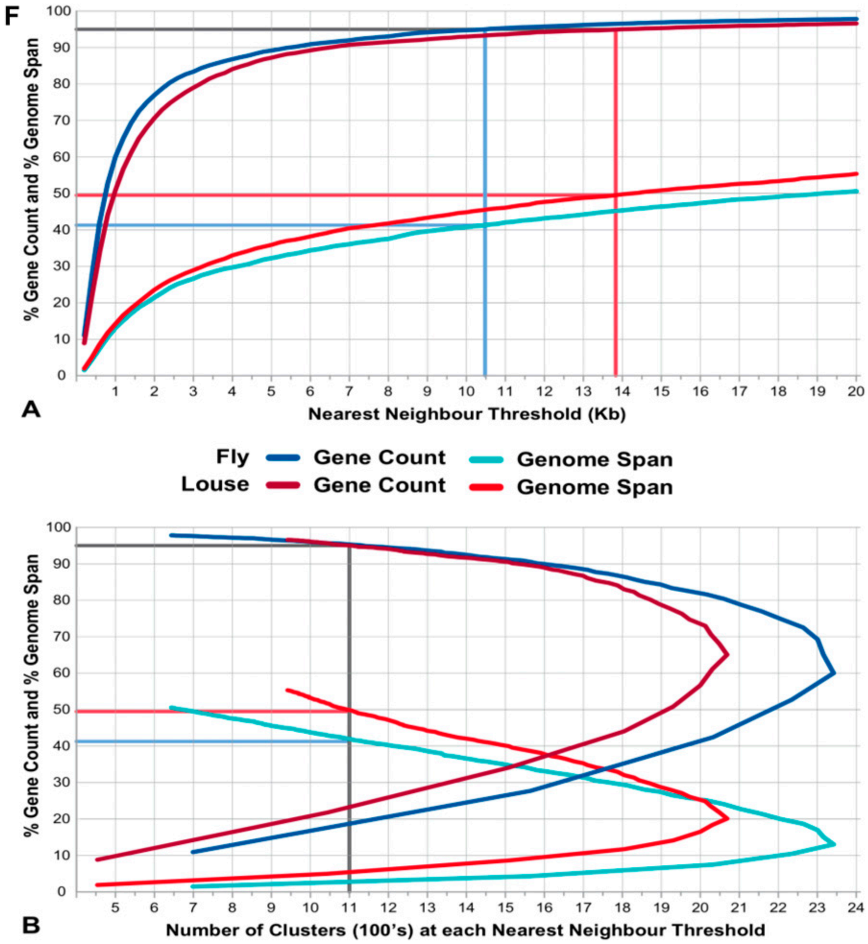


Fig. S4. Multiple alignment of microRNA genes well-represented in insect genomes and found in at least a few more basal lineages (e.g., crustaceans; shown in bold) that we failed to identify in both the *Pediculus humanus humanus* genome and raw sequencing reads; this suggests an evolutionary loss of these genes: (*A*) miR-29, (*B*) miR-33, (*C*) miR-283, and (*D*) miR-315. (*E*) Orthologous group expansions. The *P. h. humanus* (*Phum*) proteome was compared with the insects *D. melanogaster* (*Dmel*), *T. castaneum* (*Tcas*), and *Nasonia vitripennis* (Nvit) and the outgroup species *Daphnia pulex* (*Dpul*) and *Homo sapiens* (*Hsap*) to delineate groups of orthologous protein-coding genes (Fig. 1). Examining 633 expanded groups with members in all four insects reveals a lower number of expansions and significantly smaller proportions of *Phum* proteins in these expanded orthologous groups. The examined groups were required to have at least one member from each of the four insect species and a minimum of six proteins in total. These expanded groups, therefore, exhibit a minimum of a duplication in two species or a triplication in one species. Less than one-half of the groups show an expansion in *Phum* (47% > 1 member), whereas the other species exhibit more expansions (*Nvit*, 59%; *Tcas*, 70%; *Dmel*, 64%). *Phum* also shows lower mean and median values for the proportions of orthologous group members as shown in the figure box plots, and paired Wilcoxon signed-rank tests show these differences to be statistically significant. (*F*) Gene-rich portions of the *P. humanus* (louse) and *D. melanogaster* (fly) genomes. General feature format (GFF) files for louse (VectorBase PhumU1.2) and fly (FlyBase Dmel5.23) gene sets were interrogated to calculate gene spans and intergenic distances defined by protein-coding gene start and stop codons. The transcript with the longest CDS was used for genes with alternative transcripts. Merging of overlapping or intronic genes ensured that each genomic region was only counted one time in the sum of genomic spans. The numbers of genes and their total genomic spans (gene plus intergenic) were summed for intergenic thresholds in 200-bp steps up to 20 kb.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (XLS)
Table S4 (XLS)