Corrections

EVOLUTION

Correction for "Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle," by Ewen F. Kirkness, Brian J. Haas, Weilin Sun, Henk R. Braig, M. Alejandra Perotti, John M. Clark, Si Hyeock Lee, Hugh M. Robertson, Ryan C. Kennedy, Eran Elhaik, Daniel Gerlach, Evgenia V. Kriventseva, Christine G. Elsik, Dan Graur, Catherine A. Hill, Jan A. Veenstra, Brian Walenz, José Manuel C. Tubío, José M. C. Ribeiro, Julio Rozas, J. Spencer Johnston, Justin T. Reese, Aleksandar Popadic, Marta Tojo, Didier Raoult, David L. Reed, Yoshinori Tomoyasu, Emily Krause, Omprakash Mittapalli, Venu M. Margam, Hong-Mei Li, Jason M. Meyer, Reed M. Johnson, Jeanne Romero-Severson, Janice Pagel VanZee, David Alvarez-Ponce, Filipe G. Vieira, Montserrat Aguadé, Sara Guirao-Rico, Juan M. Anzola, Kyong S. Yoon, Joseph P. Strycharz, Maria F. Unger, Scott Christley, Neil F. Lobo, Manfredo J. Seufferheld, NaiKuan Wang, Gregory A. Dasch, Claudio J. Struchiner, Greg Madey, Linda I. Hannick, Shelby Bidwell, Vinita Joardar, Elisabet Caler, Renfu Shao, Stephen C. Barker, Stephen Cameron, Robert V. Bruggner, Allison Regier, Justin Johnson, Lakshmi Viswanathan, Terry R. Utterback, Granger G. Sutton, Daniel Lawson, Robert M. Waterhouse, J. Craig Venter, Robert L. Strausberg, May R. Berenbaum, Frank H. Collins, Evgeny M. Zdobnov, and Barry R. Pittendrigh, which appeared in issue 27, July 6, 2010, of *Proc Natl Acad Sci USA* (107:12168–12173; first published June 21, 2010; 10.1073/pnas.1003379107).

The authors note that the author name Emily Krause should have appeared as Emily Kraus. The corrected author line appears below. The online version has been corrected.

Ewen F. Kirkness[a,1], Brian J. Haas[a,2], Weilin Sun[b], Henk R. Braig[c], M. Alejandra Perotti[d], John M. Clark[e], Si Hyeock Lee[f], Hugh M. Robertson[b], Ryan C. Kennedy[g,h], Eran Elhaik[i], Daniel Gerlach[j,k], Evgenia V. Kriventseva[j,k], Christine G. Elsik[l,3], Dan Graur[i], Catherine A. Hill[m], Jan A. Veenstra[n], Brian Walenz[a], José Manuel C. Tubío[o], José M. C. Ribeiro[p], Julio Rozas[q], J. Spencer Johnston[r], Justin T. Reese[l], Aleksandar Popadic[s], Marta Tojo[t], Didier Raoult[u], David L. Reed[v], Yoshinori Tomoyasu[w,4], Emily Kraus[w], Omprakash Mittapalli[x], Venu M. Margam[m], Hong-Mei Li[b], Jason M. Meyer[m], Reed M. Johnson[b], Jeanne Romero-Severson[g,y], Janice Pagel VanZee[m], David Alvarez-Ponce[q], Filipe G. Vieira[q], Montserrat Aguadé[q], Sara Guirao-Rico[q], Juan M. Anzola[l], Kyong S. Yoon[e], Joseph P. Strycharz[e], Maria F. Unger[g,y], Scott Christley[g,h], Neil F. Lobo[g,y], Manfredo J. Seufferheld[z], NaiKuan Wang[aa], Gregory A. Dasch[bb], Claudio J. Struchiner[cc], Greg Madey[g,h], Linda I. Hannick[a], Shelby Bidwell[a], Vinita Joardar[a], Elisabet Caler[a], Renfu Shao[dd], Stephen C. Barker[dd], Stephen Cameron[ee], Robert V. Bruggner[g,h], Allison Regier[g,h], Justin Johnson[a], Lakshmi Viswanathan[a], Terry R. Utterback[a], Granger G. Sutton[a], Daniel Lawson[ff], Robert M. Waterhouse[j,k], J. Craig Venter[a], Robert L. Strausberg[a], May R. Berenbaum[b], Frank H. Collins[g,y], Evgeny M. Zdobnov[j,k,gg,1], and Barry R. Pittendrigh[b,1,5]

www.pnas.org/cgi/doi/10.1073/pnas.1103909108

PHYSICS

Correction for "Formation of a crystal nucleus from liquid," by Takeshi Kawasaki and Hajime Tanaka, which appeared in issue 32, August 10, 2010, of *Proc Natl Acad Sci USA* (107:14036–14041; first published July 27, 2010; 10.1073/pnas.1001040107).

The authors note that Fig. 6 appeared incorrectly. The corrected figure and its legend appear below. This error does not affect the conclusions of the article.

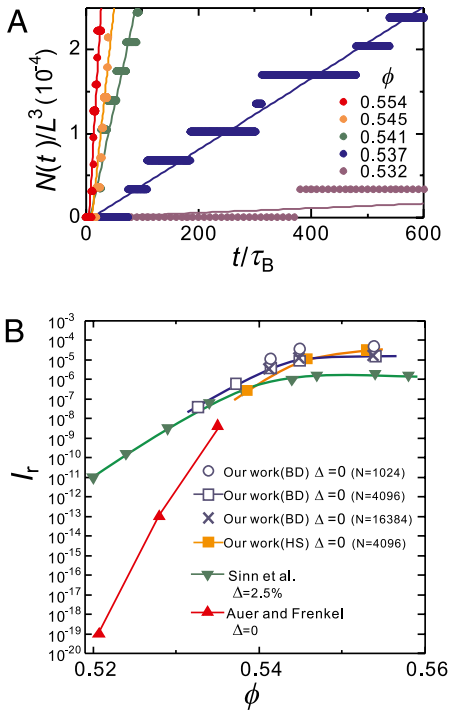

Fig. 6. Crystal nucleation dynamics. (*A*) Temporal change of the number of crystal nuclei for a system of $N = 4{,}096$ (*SI Text*). From the rate of the increase in the number of crystal nuclei, we estimated the crystal nucleation frequency I. The numbers in the figure indicate the volume fraction ϕ. (*B*) The volume fraction ϕ dependence of the reduced crystal nucleation frequency I_r for our work, the numerical estimate by Auer and Frenkel (15), and the experimental work by Sinn et al. (17). Curves are guides to the eye. We also show the results for three different system sizes (N = 1,024, 4,096, and 16,834), which indicate few finite size effects for $N \geq 4{,}096$. Here we use the volume fraction ϕ estimated with $\sigma_{eff} = 1.0953\sigma$. Here BD stands for Brownian Dynamics simulations of the WCA system and HS stands for event-driven Molecular Dynamics simulations of the hard sphere system.

www.pnas.org/cgi/doi/10.1073/pnas.1104042108

MEDICAL SCIENCES, CHEMISTRY

Correction for "Multistage nanoparticle delivery system for deep penetration into tumor tissue," by Cliff Wong, Triantafyllos Stylianopoulos, Jian Cui, John Martin, Vikash P. Chauhan, Wen Jiang, Zoran Popović, Rakesh K. Jain, Moungi G. Bawendi, and Dai Fukumura, which appeared in issue 6, February 8, 2011 of *Proc Natl Acad Sci USA* (108:2426–2431; first published January 18, 2011; 10.1073/pnas.1018382108).

The authors note that Fig. 2 and its corresponding legend appeared incorrectly. This error does not affect the conclusions of the article. The corrected figure and its corrected legend appear below.

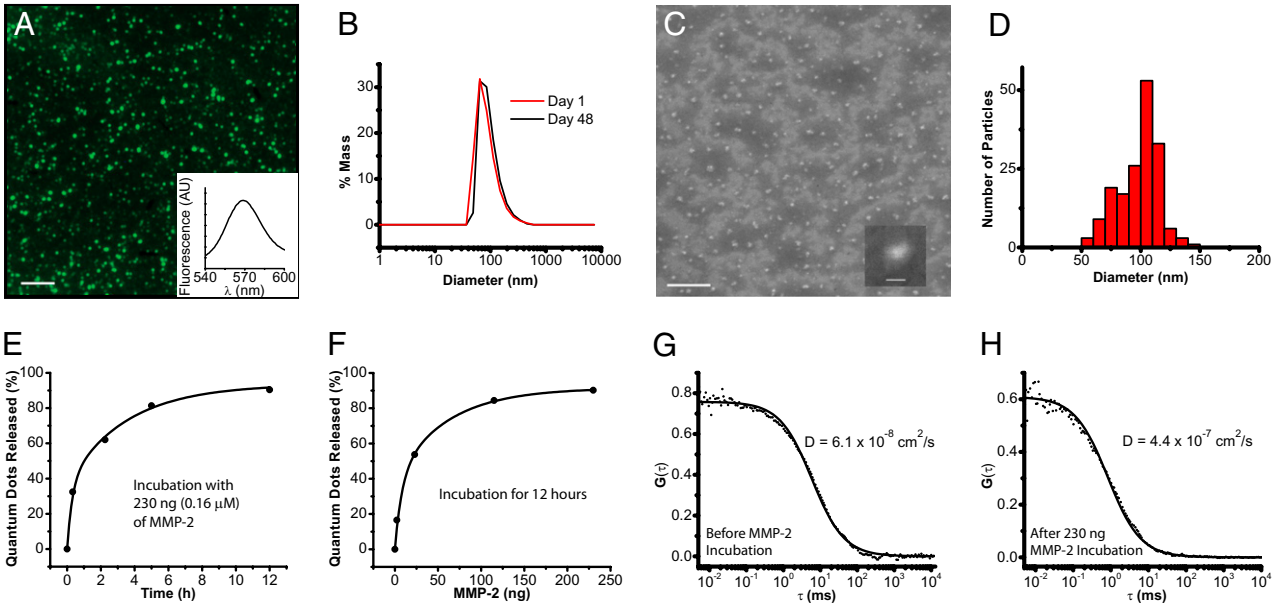

Fig. 2. QDGelNP physical and in vitro characterization. (*A*) Epifluorescence image of QDGelNPs on a silicon substrate at 100× magnification. (Scale bar: 5 μm.) (*B*) DLS distribution of QDGelNP on day 1 and day 48 after synthesis and storage at 4 °C. (*C*) SEM image of QDGelNPs at 15,000× magnification. (Scale bar: 1 μm.) (*C Inset*) SEM image of individual QDGelNP at 35,000× magnification. (Scale bar: 100 nm.) (*D*) Histogram of QDGelNPs' size distribution from image analysis of SEM image. (*E* and *F*) Kinetics of MMP-2–induced QD release from QDGelNPs. (*E*) QD-release curve from incubation of 0.1 mg of QDGelNPs with 230 ng (0.16 μM) of MMP-2. (*F*) QD release from incubation of 0.1 mg of QDGelNPs for 12 h with varying amounts of MMP-2. (*G* and *H*) FCS cross-correlograms of QDGelNPs before (*G*) and after (*H*) incubation with MMP-2.

www.pnas.org/cgi/doi/10.1073/pnas.1104327108

Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle

Ewen F. Kirkness[a,1], Brian J. Haas[a,2], Weilin Sun[b], Henk R. Braig[c], M. Alejandra Perotti[d], John M. Clark[e], Si Hyeock Lee[f], Hugh M. Robertson[b], Ryan C. Kennedy[g,h], Eran Elhaik[i], Daniel Gerlach[j,k], Evgenia V. Kriventseva[j,k], Christine G. Elsik[l,3], Dan Graur[i], Catherine A. Hill[m], Jan A. Veenstra[n], Brian Walenz[a], José Manuel C. Tubío[o,r], José M. C. Ribeiro[p], Julio Rozas[q], J. Spencer Johnston[r], Justin T. Reese[l], Aleksandar Popadic[s], Marta Tojo[t], Didier Raoult[u], David L. Reed[v], Yoshinori Tomoyasu[w,4], Emily Kraus[w], Omprakash Mittapalli[x], Venu M. Margam[m], Hong-Mei Li[b], Jason M. Meyer[m], Reed M. Johnson[b], Jeanne Romero-Severson[g,y], Janice Pagel VanZee[m], David Alvarez-Ponce[q], Filipe G. Vieira[q], Montserrat Aguadé[q], Sara Guirao-Rico[q], Juan M. Anzola[l], Kyong S. Yoon[e], Joseph P. Strycharz[e], Maria F. Unger[g,y], Scott Christley[g,h], Neil F. Lobo[g,y], Manfredo J. Seufferheld[z], NaiKuan Wang[aa], Gregory A. Dasch[bb], Claudio J. Struchiner[cc], Greg Madey[g,h], Linda I. Hannick[a], Shelby Bidwell[a], Vinita Joardar[a], Elisabet Caler[a], Renfu Shao[dd], Stephen C. Barker[dd], Stephen Cameron[ee], Robert V. Bruggner[g,h], Allison Regier[g,h], Justin Johnson[a], Lakshmi Viswanathan[a], Terry R. Utterback[a], Granger G. Sutton[a], Daniel Lawson[ff], Robert M. Waterhouse[j,k], J. Craig Venter[a], Robert L. Strausberg[a], May R. Berenbaum[b], Frank H. Collins[g,y], Evgeny M. Zdobnov[j,k,gg,1], and Barry R. Pittendrigh[b,1,5]

[a]J. Craig Venter Institute, Rockville, MD 20850; [b]Department of Entomology, University of Illinois, Urbana, IL 61801; [c]School of Biological Sciences, Bangor University, Bangor, Wales LL57 2UW, United Kingdom; [d]School of Biological Sciences, University of Reading, Reading RG6 6AS, United Kingdom; [e]Department of Veterinary and Animal Science, University of Massachusetts, Amherst, MA 01003; [f]Department of Agricultural Biotechnology, Seoul National University, Seoul, South Korea; [g]Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556; [h]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556; [i]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204; [j]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland; [k]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland; [l]Department of Animal Science, Texas A&M University, College Station, TX 77843; [m]Department of Entomology, Purdue University, West Lafayette, IN 47907; [n]Centre National de la Recherche Scientifique, Centre Neurosciences Intégratives et Cognitives, University of Bordeaux, 33405 Talence Cedex, France; [o]Servicio de Hematoloxía, Complexo Hospitalario Universitario de Santiago de Compostela, 15706 Santiago de Compostela, Spain; [p]Laboratory of Malaria and Vector Research, National Institutes of Health, Bethesda, MD 20892-8132; [q]Departament de Genètica, Universitat de Barcelona, 08028 Barcelona, Spain; [r]Department of Entomology, Texas A&M University, College Station, TX 77843; [s]Department of Biological Sciences, Wayne State University, Detroit, MI 48202; [t]Servicio de Anatomía Patolóxica, Complexo Hospitalario Universitario de Santiago de Compostela, 15706 Santiago de Compostela, Spain; [u]Unité des Rickettsies, 13385 Marseille Cedex 05, France; [v]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; [w]Division of Biology and Department of Entomology, Kansas State University, Manhattan, KS 66502; [x]Department of Entomology, Ohio Agriculture Research and Development Center/Ohio State University, Wooster, OH 44691; [y]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; [z]Department of Crop Sciences, Urbana, IL 61801; [aa]Chung Hwa College of Medical Technology, Jen-Te Hsiang, Tainan 700, Taiwan; [bb]Centers for Disease Control and Prevention, Atlanta, GA 30333; [cc]Escola Nacional de Saúde Pública Sergio Arouca/Fundação Oswaldo Cruz and Instituto de Medicina Social Universidade do Estado do Rio de Janeiro, 4365 Rio de Janeiro, Brazil; [dd]School of Chemistry and Molecular Biosciences, University of Queensland, St. Lucia, Brisbane, Queensland 4072, Australia; [ee]Australian National Insect Collection and Commonwealth Scientific and Industrial Research Organization Entomology, Canberra ACT 2601, Australia; [ff]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom; and [gg]Imperial College London, London SW7 2AZ, United Kingdom.

Contributed by May R. Berenbaum, April 14, 2010 (sent for review February 10, 2010)

As an obligatory parasite of humans, the body louse (*Pediculus humanus humanus*) is an important vector for human diseases, including epidemic typhus, relapsing fever, and trench fever. Here, we present genome sequences of the body louse and its primary bacterial endosymbiont *Candidatus* Riesia pediculicola. The body louse has the smallest known insect genome, spanning 108 Mb. Despite its status as an obligate parasite, it retains a remarkably complete basal insect repertoire of 10,773 protein-coding genes and 57 microRNAs. Representing hemimetabolous insects, the genome of the body louse thus provides a reference for studies of holometabolous insects. Compared with other insect genomes, the body louse genome contains significantly fewer genes associated with environmental sensing and response, including odorant and gustatory receptors and detoxifying enzymes. The unique architecture of the 18 minicircular mitochondrial chromosomes of the body louse may be linked to the loss of the gene encoding the mitochondrial single-stranded DNA binding protein. The genome of the obligatory louse endosymbiont *Candidatus* Riesia pediculicola encodes less than 600 genes on a short, linear chromosome and a circular plasmid. The plasmid harbors a unique arrangement of genes required for the synthesis of pantothenate, an essential vitamin deficient in the louse diet. The human body louse, its primary endosymbiont, and the bacterial pathogens that it vectors all possess genomes reduced in size compared with their free-living close relatives. Thus, the body louse genome project offers unique information and tools to use in advancing understanding of coevolution among vectors, symbionts, and pathogens.

Author contributions: E.F.K., B.J.H., J.M.C., H.M.R., R.C.K., E.E., C.G.E., D. Graur, C.A.H., B.W., J.M.C.T., J.M.C.R., J.R., J.S.J., J.T.R., M.T., D.L.R., O.M., V.M.M., J.M.M., J.R.-S., J.P.V., M.A., J.M.A., K.S.Y., J.P.S., M.F.U., S. Christley, N.F.L., G.A.D., C.J.S., G.M., S. Cameron, A.R., G.G.S., J.C.V., R.L.S., F.H.C., E.M.Z., and B.R.P. designed research; B.J.H., W.S., J.M.C., S.H.L., R.C.K., E.E., D. Gerlach, E.V.K., C.G.E., D. Graur, C.A.H., J.A.V., B.W., J.M.C.R., J.R., J.S.J., J.T.R., A.P., Y.T., M.T., D.R., D.L.R., E.K., O.M., H.-M.L., J.M.M., R.J., J.R.-S., J.P.V., D.A.-P., F.G.V., M.A., S.G.-R., J.M.A., K.S.Y., J.P.S., M.F.U., S. Christley, M.J.S., N.W., C.J.S., R.S., S.C.B., A.R., J.J., L.V., T.R.U., D.L., R.M.W., M.R.B., E.M.Z., and B.R.P. performed research; J.M.C., R.C.K., E.E., C.G.E., D. Graur, C.A.H., J.M.C.T., J.S.J., J.T.R., M.T., J.M.M., J.R.-S., J.M.A., K.S.Y., J.P.S., M.F.U., S. Christley, L.I.H., V.J., E.C., R.S., S.C.B., R.V.B., L.V., R.M.W., and E.M.Z., contributed new reagents or analytic tools; E.F.K., W.S., H.R.B., M.A.P., S.H.L., H.M.R., R.C.K., D. Gerlach, E.V.K., J.A.V., B.W., J.M.C.T., J.M.C.R., J.R., A.P., Y.T., M.T., D.L.R., H.-M.L., R.J., J.R.-S., D.A.-P., F.G.V., M.A., S.G.-R., K.S.Y., J.P.S., M.F.U., S. Christley, M.J.S., N.W., G.A.D., C.J.S., L.I.H., S.C.B., V.J., E.C., R.S., S. Cameron, R.V.B., A.R., J.J., D.L., R.M.W., M.R.B., F.H.C., E.M.Z., and B.R.P. analyzed data; and E.F.K., B.J.H., W.S., H.R.B., M.A.P., S.H.L., H.M.R., R.C.K., B.W., J.M.C.T., J.R., A.P., Y.T., M.T., D.R., D.L.R., O.M., R.J., D.A.-P., F.G.V., M.A., S.G.-R., J.M.A., K.S.Y., J.P.S., M.F.U., S. Christley, G.M., S. Cameron, A.R., R.M.W., M.R.B., F.H.C., E.M.Z., and B.R.P. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AAZO00000000, NC_014109, and NC_013962).

[1]E.F.K., E.M.Z., and B.R.P. contributed equally to this work.

[2]Present address: Broad Institute, 7 Cambridge Center, MA 02142.

[3]Present address: Department of Biology, Georgetown University, Washington, DC 20057.

[4]Present address: Department of Zoology, Miami University, 252 Pearson Hall, Oxford, OH 45056.

[5]To whom correspondence should be addressed. E-mail: pittendr@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1003379107/-/DCSupplemental.

ectoparasite | comparative genomics | coevolution

Like their primate relatives, humans have had a long evolutionary association with parasitic sucking lice. Contact between sucking lice and primate hosts dates back at least 25 million years (1). Chimpanzee lice (*Pediculus schaeffi*) and human lice (*Pediculus humanus*) diverged from their common ancestors, as did chimpanzees (*Pan troglodytes*) and humans (*Homo sapiens*), 5–7 million years ago (2, 3). The two subspecies—the human body louse (*Pediculus humanus humanus* L.) and the head louse (*P. h. capitis* DG.)—are closely related obligate parasites that feed exclusively on human blood. Body lice likely evolved from head louse ancestors when humans began to wear clothing, which is required for egg deposition by body lice (4).

P. h. humanus has been of tremendous medical and social importance throughout human history. Of the two forms, only the body louse has been implicated as a vector of human disease and is the principal vector of epidemic typhus (*Rickettsia prowazekii*), relapsing fever (*Borrelia recurrentis*), and trench fever (*Bartonella quintana*) (5–9). In the United States as well as the rest of the world, body lice are primarily a concern in transient homeless populations, whereas head lice tend to infest populations of elementary school-aged children. Historically, epidemic typhus has been responsible for massive mortality in wartime (9); in contemporary times, major outbreaks of epidemic typhus are found primarily among refugees [e.g., in Burundi in 1996 (8)], but sporadic cases have also been observed in general populations in Russia (10), Peru, Algeria, and France (11).

Like all hematophagous lice, body lice depend on obligate endosymbionts to supplement their nutritionally deficient blood diet (12). The primary endosymbiont of *P. h. humanus* has been given the provisional name *Candidatus* Riesia pediculicola (13) (hereafter, Riesia). The body louse maintains organs called mycetomes that house the primary endosymbiont, except during passage to the ovaries for transovarial transmission (14). The tripartite interdependency of this bacterial endosymbiont, its body louse host, and the human host of the body louse seems to have coevolved over several million years (15).

Here, we present the genome sequences of the body louse and its coevolved primary endosymbiont. This genome, the smallest known insect genome, encodes a remarkably complete gene repertoire and thus, provides a robust phylogenetic outgroup for understanding the evolution of holometabolous insects. The striking reduction in genome size is particularly notable in gene families associated with environmental sensing and response; this reduction befits a monophagous permanent parasite with a substantially reduced need to seek out food sources and detect and avoid enemies relative to free-living species.

Results and Discussion

Genome Features. ***Genome sequencing, assembly, and annotation.*** The genome of the body louse was sequenced to 8.5× average coverage using a whole genome shotgun approach with 1.3 million paired-end reads from plasmid libraries. The assembled contigs and scaffolds, spanning 108 Mb and 110 Mb, respectively, confirmed previous estimates based on flow cytometry data (103–109 Mb) that the body louse has the smallest known genome size of any insect (16, 17). The 300 longest scaffolds span more than 95% of the assembled genome sequence (scaffold N50 size of 488 kb). A range of automated and manual methods (18) yielded 10 tentative superscaffolds of up to 9 Mb each, spanning a total of 49 Mb. This effort provided large chromosomal segments, which were close to continuous, with only a few remaining clone gaps, usually involving simple-sequence gene deserts.

The remarkable compactness of the genome greatly facilitated accurate gene annotation. Predictions using multiple gene-modeling approaches resulted in consensus annotation (Table 1) of 10,773 protein-coding genes, 161 transfer ribonucleic acids (tRNAs) for all 20 amino acids, and 57 microRNAs (Table S1*A*). Comparing predicted protein lengths with their *Drosophila melanogaster* orthologs (the best experimentally studied insect that drives comparative gene annotation) revealed greater consistency with body louse genes (concordance = 0.91; identical with *Anopheles gambiae*) than with the honey bee *Apis mellifera* (concordance = 0.89) or the red flour beetle *Tribolium castaneum* (concordance = 0.88), despite greater evolutionary divergence (Fig. S1*A*).

GC content. Compared with other sequenced insect genomes, the body louse genome has the highest abundance of small homogeneous GC-content domains (7–30 kb with GC content between 18% and 63%). The average GC content of the *P. h. humanus* genome is 28%, which is similar to that of the *A. mellifera* genome (33%), making these two genomes unusually AT-rich. However, the *A. mellifera* genome harbors more extremes. Only 77% of homogeneous domains have a GC content between 20% and 60% in *A. mellifera* compared with 94% in *P. h. humanus*, which is more similar in this respect to the genome of *T. castaneum* (99%) (Fig. S2 *A* and *B*).

Telomeres. Unlike *A. mellifera* telomeres (19), none of the body louse telomeres appeared to be assembled completely at the ends of long superscaffolds. Therefore, we sought candidate telomere sequences with the strategy used for *T. castaneum* (20). The body louse is diploid, and it has a haploid complement of five metacentric chromosomes and one telocentric chromosome for a total of 11 putative telomeres (21). Although we were unable to reconstruct an entire telomere because of its highly repetitive nature, we identified a long subtelomeric repeat region that was partially assembled on at least 9 of 11 putative telo-

Table 1. Summary of the genome features of *Pediculus humanus humanus* compared with *Drosophila melanogaster*

Genome feature	Count	Nucleotides (Mb)	Genome fraction (%)
P. h. humanus (*D. melanogaster*)	6 chromosomes (4 chromosomes)	110 (169)	100 (100)
Gene-rich clusters* containing 95% of genes	1,110 (1,130)	55 (70)	50 (41)
Protein-coding genes			
Total [multi-exon]	10,773, [10,424]; (13,794, [11,458])	33.8 (82.6)	31 (49)
Coding exons	69,261 (54,606)	16.6 (22.3)	15 (13)
Introns	58,522 (44,698)	17.2 (48.6)	15 (29)
Non–protein-coding genes			
tRNAs	161 (292)	0.012 (0.022)	<1
miRNAs	57 (90)	0.005 (0.008)	<1
Transposable elements	3,558 (9,409)	1.1 (11.6)	1 (7)
Tandem repeats	130,608 (25,904)	6.9 (6.1)	6 (4)

D. melanogaster values were obtained from FlyBase release 5.23 with the same parameters used to obtain, parse, and count the *P. h. humanus* genome. The more numerous body louse exons and introns suggest intron loss in *D. melanogaster* but with an increase in their sizes.
*Supporting documentation is in Fig. S4*F*.

meres between unique flanking DNA and telomeric TTAGG repeats. This subtelomeric region consists of various satellite-like repeats in addition to pseudogenes and simple sequences, and it varies considerably in length. The TTAGG repeats commonly contain sequence associated repeat telomeric (SART)-like retrotransposons, which are also characteristic of the telomeres from *T. castaneum* and *Bombyx mori* (domestic silkworm). This combination might represent the basal insect situation. If so, the simple TTAGG telomeres of *A. mellifera* would represent a derived condition in which most retrotransposons have been lost rather than the ancestral condition (19). Alternatively, insect telomeres may have repeatedly been invaded as a safe harbor by non-LTR retrotransposons of the R-element family that belongs to the SART group (20).

Transposable elements. Both class I and class II mobile elements are present in the genome of *P. h. humanus*, yet they represent only 1% of the genome (Table S1*B*), which is markedly lower than any sequenced insect genome. Interestingly, the body louse genome size is near the hypothesized 100 Mb critical threshold at which transposable elements can be established in eukaryote genomes (22).

Mitochondrial genome. The mitochondrial genome of *P. h. humanus* contains the full complement of 37 genes organized in an unusual architecture of 18 minicircular chromosomes (23). It is possible that multiple minicircular chromosomes promote recombination between genes on different chromosomes. Indeed, there is evidence in the genome sequence data for at least two chimeric minicircular chromosomes that have arisen from such recombination (Fig. S1*B*).

Of 305 mitochondrial-targeted, nuclear-encoded genes known in *D. melanogaster,* 282 have louse orthologs. This finding suggests that the basic mitochondrial functions (e.g., oxidative phosphorylation, membrane transport, and protein synthesis) are unimpeded by the reorganized mitochondrial genome. The body louse genome revealed the apparent loss of the mitochondrial single-stranded binding protein (mtSSB), a factor required for optimal initiation and processivity during mitochondrial genome replication in both insects and mammals (24, 25). In the absence of mtSSB, complete replication of a full-sized mitochondrial genome may not be possible (25); the loss of mtSSB function in *D. melanogaster* is lethal at the late third instar/pupal stages because of a loss of mtDNA content (26). It is not yet known if the mtSSB function can be replaced by an endosymbiont homolog or if the multiple minicircles render the mtSSB unnecessary.

Endosymbiont Genome. ***Genome sequencing, assembly, and annotation.*** Like many other sucking lice (Anoplura, Rhyncophthirina), the body louse has mycetomes that harbor the primary endosymbiotic bacteria (p-endosymbionts). The genome of the *Pediculus* symbiont, Riesia, was sequenced to an average coverage of 50× and is composed of a single linear chromosome of at least 574,526 bp with palindromic termini and a single circular plasmid of 7,628 bp. The chromosome contains 557 ORFs, 33 tRNAs, 6 ribosomal RNAs, and 1 other structural RNA.

Comparisons with other endosymbionts. We compared the genome of Riesia with the genomes of other endosymbionts and the infectious plague pathogen *Yersinia pestis* (Fig. S3). This genome-wide sequence comparison revealed a core of 237 genes common to all bacteria examined; only 24 genes were unique to Riesia, and 30 genes were present in all except Riesia (Table S2 *A* and *B*). Several genes unique to Riesia code for transport and binding proteins as well as for enzymes involved in lipopolysaccharide biosynthesis. Conversely, the enzymes missing from Riesia are mainly exonucleases, which are required for conjugation, and enzymes involved in energy metabolism. The Riesia-specific transport and binding proteins and the lack of energy metabolism genes may reflect the dependence of Riesia on its louse host for nutrients. Lipopolysaccharides might be important for cell-wall stability when Riesia migrate extracellularly through the louse to reach filial mycetomes in the ovaries (14) (Table S2*B*).

Riesia is required by lice for the production of pantothenic acid (vitamin B5). Without Riesia, nymphs die during their first molt (27). Surprisingly, the genes for three key enzymes in the synthesis of pantothenic acid, *panB*, *panC*, and *panE*, are missing from the linear chromosome of Riesia. These genes are, instead, found together on the plasmid. Similar cases are known from evolutionarily more ancient endosymbionts (e.g., *Buchnera*) in which essential genes are also extrachromosomal (28). Having these genes on a multicopy plasmid could represent a mechanism that reduces the risk of genome degradation and increases expression levels to secure synthesis of pantothenic acid at required amounts. Interestingly, there is preliminary evidence that endosymbiont replacement may be commonplace in sucking lice (29), possibly facilitated by the acquisition of plasmids that harbor genes essential to the host.

Nakabachi et al. (30) proposed that integration of essential genes from the p-endosymbiont into the host genome might be an important mechanism for the host to overcome the consequences of genome degradation of its endosymbiont. Riesia in the human body louse and *Buchnera* in the pea aphid (*Acyrthosiphon pisum*) represent cases where the genomes of both symbiotic partners are available to test this hypothesis. The body louse genome does not appear to contain any genes of prokaryotic origin, suggesting the absence of transfers from Riesia. In the pea aphid, there is also no gene transfer from the endosymbiont, but there is evidence of gene transfer from other bacteria (31).

The dramatic reduction in genome size and high AT bias suggest a long association between Riesia and its host insect, and like some other ancient gammaproteobacterial symbiotic associations, the Riesia genome is free of mobile elements. However, Riesia's association with its host is only 13–25 million years old, making Riesia one of the youngest known endosymbionts (31).

Comparative Genomics. ***Hemimetabolous outgroup.*** The human body louse is among the first sequenced representatives of hemimetabolous insects (32), a group distinguished by progressive intermediate development as nymphal instars rather than larva–pupa–adult transformations. The louse genome is, therefore, an important outgroup reference for comparative analyses of sequenced holometabolous insects (Fig. 1*A*). The complete metamorphosis of holometabolous insects is a highly successful evolutionary strategy, whereby larvae and adults can take advantage of different ecological niches. The molecular innovations that have contributed to the success of holometabolous insects can now be viewed in the context of a hemimetabolous outgroup genome sequence that is largely complete.

In addition to being the smallest genome of any insect studied to date, the body louse genome is, as far as can be determined, functionally complete. Of 10,773 body louse protein-coding genes, 90% share homology to genes known in other species, enabling orthology delineation for 80% of louse genes (33). This level is comparable with results from initial analyses from *A. mellifera* (34) and *T. castaneum* (20). The phylogenetic tree reconstructed using single-copy orthologs (Fig. 1*A*) confirms the basal position of Hemimetabola compared with Holometabola within Insecta. This suggests an average rate of molecular evolution in the lineage of lice that is comparable with that of Hymenoptera and Coleoptera.

Microsynteny analysis (35) between genomes of the body louse and hymenopteran honey bee *A. mellifera* or *Nasonia* parasitoid wasp species suggests that about 20% of single-copy orthologs are retained in their ancestral arrangements (Table S3*A*). This percentage is similar to microsynteny conservation levels between *A. mellifera* (Hymenoptera) and *T. castaneum* (Coleoptera), and it is substantially greater than their conservation with dipterans (<15%) (36), highlighting the derived state of Diptera.

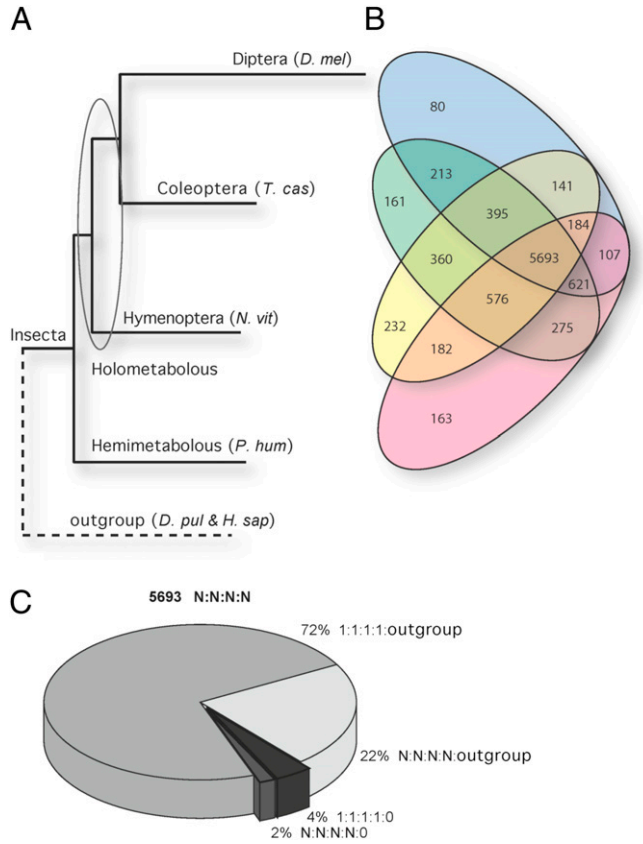

Fig. 1. The *Pediculus humanus humanus* (*P. hum*) genome reveals a basal insect gene repertoire. The encoded *P. hum* proteome is compared with sequenced representatives of the orders Diptera, Coleoptera, and Hymenoptera and outgroup species beyond Insecta. *D. mel*, *Drosophila melanogaster*; *T. cas*, *Tribolium castaneum*; *N. vit*, *Nasonia vitripennis*; *D. pul*, *Daphnia pulex; H. sap, Homo sapiens*. (*A*) The Maximum-Likelihood phylogenetic tree was reconstructed using the superalignment of protein sequences of universal single-copy orthologs. The obtained tree confirms the basal position of Hemimetabola compared with Holometabola within Insecta. The branch lengths are proportional to the accumulated number of substitutions, suggesting an average rate of molecular evolution in lice that is comparable with that in Hymenoptera and Coleoptera. (*B*) The Venn diagram shows the numbers of orthologous groups of genes shared among the four insects (a lower estimate of the ancestral number of genes). It depicts the phylogenetic distribution of orthologs, highlighting the completeness of the gene repertoire encoded in the body louse genome. Pink, *P. hum*; yellow, *N. vit*; green, *T. cas*; blue, *D. mel*. (*C*) The pie chart partitions the largest fraction of core body louse proteins with orthologs in three holometabolous insect orders and the outgroup species beyond Insecta with respect to single- (1:1:1:1) and multiple- (N:N:N:N) copy orthologs. Of 5,693 groups of single- and multiple-copy orthologs common across Insecta, 94% are shared across Bilateria as single-copy (72%) or multiple-copy (22%) orthologs, and only 6% are insect-specific orthologous groups (4% as single copies and 2% as multiple copies).

Ancestral insect gene repertoire. Contrary to the expectations of reductive evolution common in obligate parasites, the body louse has retained a remarkably complete repertoire of both protein-coding and non–protein-coding genes (Table 1). The distribution of orthologous genes across four representative insect species (Fig. 1 *B* and *C*) shows that Hymenoptera and Coleoptera share more orthologs with the body louse than they do with the fruit fly *D. melanogaster*. Relative to the well-studied *D. melanogaster* model, the louse genome may be used as a robust outgroup to Holometabola.

Examining microRNA gene families shared among crustaceans and insects revealed that mir-315, mir-283, mir-33, and mir-29 were lost from the body louse genome (37) (Table S1*A* and Fig. S4 *A–D*) (mir-iab-4 and mir-46 have been found in the trace archive). Because all true lice are wingless, it is intriguing to note that mir-315 has been identified as a potent activator of wingless signaling in *D. melanogaster* (38).

Evolution of Gene Families in Relation to the Life History of the Body Louse. The body louse has maintained many genes important for basic physiological processes, losing only a few of these roles to its endosymbiont Riesia. Because the expansion and contraction of gene families may indicate functional adaptation and evolution, we compared the body louse gene repertoire with those of the honey bee and red flour beetle. Comparisons were made both at the level of protein families, which could be generally defined using InterPro domain signatures (Table S3 *B–D*), and at a finer scale at the level of orthologous groups of genes (Fig. S4*E*). On both scales, the body louse genome seems to have several gene families with fewer members than those found in other invertebrates.

Fewer Genes Are Associated with Environmental Sensing and Response.
G protein-coupled receptors. With 104 nonsensory G protein-coupled receptors (GPCRs) and 3 opsins (visual receptors) (Table S4), *P. h. humanus* has the smallest repertoire of GPCRs identified in any sequenced insect genome to date (20, 34, 39–41). The louse genome has orthologs for ∼80% of nonsensory GPCRs identified in *D. melanogaster*. These GPCRs seemingly represent a minimal suite of receptors needed to maintain conserved GPCR-mediated signaling pathways common to diverse insect taxa (42). The relatively small number of louse opsins likely reflects its simple visual system. Moreover, the body louse lacks a putative short (blue)-wavelength sensitive opsin typically found in other insects (43), a feature that might have evolved during its adaptation to the obligate parasitic lifestyle.

Odorant-, gustatory-, and chemosensory-related genes. The genome sequence revealed just 10 odorant receptor (Or) genes, fewer than any other insect examined to date by almost an order of magnitude. The gustatory receptor (Gr) family is comparably small with just six loci encoding eight proteins through alternative splicing of the N terminus of one locus. There are no orthologs of the otherwise highly conserved carbon dioxide heterodimer Gr receptors (40, 41, 44, 45) or the putative sugar receptors (46, 47). *P. h. humanus* contains five and seven putative functional odorant-binding proteins (OBPs) and chemosensory proteins (CSPs), respectively (Table S4), and this number is dramatically less than that found in other insects (48). These aforementioned sensory genes and their resultant proteins are presumably not necessary for host location and selection. Furthermore, lice do not need to avoid the many bitter xenobiotic toxins to which most insect Grs seem to be tuned (46).

Insulin/Target of Rapamycin (TOR) pathway genes. The insulin/TOR signal transduction pathway plays a central role in multiple and critical biological processes, including organismal growth, anabolic metabolism, cell survival, fertility, and lifespan determination (49, 50). This pathway has been well-characterized in multiple organisms, including *D. melanogaster* (51). Both the structure of the pathway and the molecular function of its components are well-conserved across metazoans. The body louse genome encodes a complete insulin/TOR signaling pathway. However, these genes are reduced in number in the body louse in contrast with *D. melanogaster*, where some genes have multiple copies (Table S4*D*). Remarkably, the louse has a single insulin-like peptide (*ilp*) gene. Given that there is some evidence for differential expression of *ilp* genes under different dietary conditions in insects (52, 53), the presence of a single *ilp* gene in the body louse genome might reflect its restricted and homogeneous diet.

Detoxification enzymes. The louse genome encodes the smallest number of detoxification enzymes observed in any insect, reflecting its obligate parasite lifestyle in which it is sheltered from xenobiotic challenges faced by free-living insects (e.g., plant

secondary compounds). There are notably few cytochrome P450s and only 12 genes within the *CYP3* clade which is closely associated with xenobiotic metabolism. In contrast, *D. melanogaster* and *A. mellifera* have 36 and 28 *CYP3* clade genes, respectively. Among the 13 glutathione-S-transferases (GST) (Table S4*E*), none belong to the Epsilon class that has been shown to contribute to insect adaptation to environmental selection pressures (54). The Epsilon class was also missing in the pea aphid genome. In contrast, the relative abundance of Delta class GSTs (more than *A. mellifera*) suggests that *P. h. humanus* still possesses some capacity for detoxification of xenobiotics, including insecticides (55).

Body louse coevolution and allopatric speciation. With their characteristic extreme host specificity, pediculid lice provide dramatic examples of host–parasite coevolution and allopatric speciation (56). One consequence of this specificity is the difficulty encountered when adapting human lice to novel experimental hosts (8). Body lice have reduced genomes and harbor specific bacterial symbionts and pathogens that also exhibit genome reduction (57–64). These combined observations support the hypothesis that *P. h. humanus* has become highly specialized since its divergence from the chimpanzee louse 5–7 million years ago. Such extreme specializations in the endosymbiont, associated with dramatic genome reductions, may have resulted from a lack of gene exchange after allopatric speciation. This association of an insect host, its symbionts, and its bacterial pathogens coevolving and showing congruent reductive genome evolution provides a dramatic example of the evolutionary consequences of genome interactions and interdependency over time.

Conclusions

The body louse genome provides a unique repository of data that has considerable basic and practical significance. The availability of sequence data will facilitate molecular studies of a vector for diseases that continue to afflict human populations around the world. The louse relies on Riesia, an obligatory louse bacterial endosymbiont that lacks antibiotic resistance genes, for survival; thus, the development of louse-control strategies targeting this symbiont may be possible. With respect to understanding the evolution of multigene families mediating responses to environmental selective forces, the body louse genome, with its drastically reduced inventories in the context of its exceptionally homogeneous environment, provides extraordinary prospects for characterizing the functionalities of these rapidly evolving proteins. As well, further studies focusing on the smaller repertoire of detoxification genes and olfactory receptors in the body louse may guide the development of pediculicides and repellents with negligible impacts on human hosts. Moreover, the remarkable completeness of this genome, despite its small size, will serve as a key evolutionary reference point for studies of all sequenced insect species in characterizing the fundamental prerequisites for insect growth and development. Finally, the body louse genome will provide an opportunity for the scientific community to gain greater insights into host–parasite–symbiont tripartite coevolution and speciation.

Materials and Methods

Lice were obtained from an inbred colony derived from the Culpepper strain (65) that has been maintained on rabbits since 1999 at the University of Massachusetts, Amherst, MA. Total DNA was extracted from ~100 first instar nymphs before their first blood meal and was used to construct libraries in the plasmid, pHOS2 (3- to 4-kb and 10- to 12-kb inserts), or the fosmid, pCCFOS1 (35- to 40-kb inserts). End sequencing of clones from each library was conducted using a standard capillary platform (ABI 3730), and it yielded 1.30 million good traces (96% paired) with a mean clear read length of 656 bases. All traces were deposited in the National Center for Biotechnology Information (NCBI) trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?). The reads were assembled with Celera Assembler (http://wgs-assembler.sourceforge.net) (66–68) and deposited with NCBI (accession no. AAZO00000000). The details of the assembly and annotation are given in *SI Text*. Additional analyses of other aspects of the body louse genome are given in the *SI Text*.

ACKNOWLEDGMENTS. We thank Georg Jander and Marian Goldsmith for helpful discussion and manuscript review.

1. Reed DL, Light JE, Allen JM, Kirchman JJ (2007) Pair of lice lost or parasites regained: The evolutionary history of anthropoid primate lice. *BMC Biol* 5:7.
2. Reed DL, Smith VS, Hammond SL, Rogers AR, Clayton DH (2004) Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS Biol* 2:e340.
3. Light JE, Reed DL (2009) Multigene analysis of phylogenetic relationships and divergence times of primate sucking lice (Phthiraptera: Anoplura). *Mol Phylogenet Evol* 50:376–390.
4. Kittler R, Kayser M, Stoneking M (2003) Molecular evolution of *Pediculus humanus* and the origin of clothing. *Curr Biol* 13:1414–1417.
5. Eremeeva ME, Madan A, Shaw CD, Tang K, Dasch GA (2005) New perspectives on rickettsial evolution from new genome sequences of *Rickettsia*, particularly *R. canadensis*, and *Orientia tsutsugamushi*. *Ann N Y Acad Sci* 1063:47–63.
6. Rotz LD, Khan AS, Lillibridge SR, Ostroff SM, Hughes JM (2002) Public health assessment of potential biological terrorism agents. *Emerg Infect Dis* 8:225–230.
7. Andersson JO, Andersson SG (2000) A century of typhus, lice and *Rickettsia*. *Res Microbiol* 151:143–150.
8. Raoult D, Roux V (1999) The body louse as a vector of reemerging human diseases. *Clin Infect Dis* 29:888–911.
9. Raoult D, et al. (2006) Evidence for louse-transmitted diseases in soldiers of Napoleon's Grand Army in Vilnius. *J Infect Dis* 193:112–120.
10. Tarasevich I, Rydkina E, Raoult D (1998) Outbreak of epidemic typhus in Russia. *Lancet* 352:1151.
11. Bechah Y, Capo C, Mege JL, Raoult D (2008) Epidemic typhus. *Lancet Infect Dis* 8: 417–426.
12. Buchner P (1965) *Endosymbiosis of Animals with Plant Microorganisms* (Interscience Publishers, New York), pp 909.
13. Sasaki-Fukatsu K, et al. (2006) Symbiotic bacteria associated with stomach discs of human lice. *Appl Environ Microbiol* 72:7349–7352.
14. Perotti MA, Allen JM, Reed DL, Braig HR (2007) Host-symbiont interactions of the primary endosymbiont of human head and body lice. *FASEB J* 21:1058–1066.
15. Allen JM, Reed DL, Perotti MA, Braig HR (2007) Evolutionary relationships of "*Candidatus* Riesia spp.," endosymbiotic Enterobacteriaceae living within hematophagous primate lice. *Appl Environ Microbiol* 73:1659–1664.
16. Pittendrigh BR, et al. (2006) Sequencing of a new target genome: The *Pediculus humanus humanus* (Phthiraptera: Pediculidae) genome project. *J Med Entomol* 43: 1103–1111.
17. Johnston JS, Yoon KS, Strycharz JP, Pittendrigh BR, Clark JM (2007) Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. *J Med Entomol* 44:1009–1012.
18. Robertson HM, et al. (2007) Manual superscaffolding of honey bee (*Apis mellifera*) chromosomes 12-16: Implications for the draft genome assembly version 4, gene annotation, and chromosome structure. *Insect Mol Biol* 16:401–410.
19. Robertson HM, Gordon KH (2006) Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*. *Genome Res* 16:1345–1351.
20. Richards S, et al. (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452:949–955.
21. Hindle E, Pontecorvo G (1942) Mitotic divisions following meiosis in *Pediculus corporis* males. *Nature* 149:668.
22. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404.
23. Shao R, Kirkness EF, Barker SC (2009) The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome Res* 19:904–912.
24. Farr CL, Matsushima Y, Lagina AT, 3rd, Luo N, Kaguni LS (2004) Physiological and biochemical defects in functional interactions of mitochondrial DNA polymerase and DNA-binding mutants of single-stranded DNA-binding protein. *J Biol Chem* 279: 17047–17053.
25. Korhonen JA, Pham XH, Pellegrini M, Falkenberg M (2004) Reconstitution of a minimal mtDNA replisome *in vitro*. *EMBO J* 23:2423–2429.
26. Maier D, et al. (2001) Mitochondrial single-stranded DNA-binding protein is required for mitochondrial DNA replication and development in *Drosophila melanogaster*. *Mol Biol Cell* 12:821–830.
27. Perotti MA, Kirkness EF, Reed DL, Braig HR (2009) Endosymbionts of lice. *Insect Symbiosis 3*, ed Bourtzis KMT (Taylor & Francis, Boca Raton, FL), pp 205–220.
28. Ding H, Hynes MF (2009) Plasmid transfer systems in the rhizobia. *Can J Microbiol* 55: 917–927.
29. Hypsa V, Krizek J (2007) Molecular evidence for polyphyletic origin of the primary symbionts of sucking lice (phthiraptera, anoplura). *Microb Ecol* 54:242–251.
30. Nakabachi A, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267.
31. Allen JM, Light JE, Perotti MA, Braig HR, Reed DL (2009) Mutational meltdown in primary endosymbionts: Selection limits Muller's ratchet. *PLoS One* 4:e4969.
32. The International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* 8:e1000313.

33. Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: The hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* 36:D271–D275.
34. Weinstock GM, et al. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.
35. Zdobnov EM, et al. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298:149–159.
36. Zdobnov EM, Bork P (2007) Quantification of insect genome divergence. *Trends Genet* 23:16–20.
37. Gerlach D, Kriventseva EV, Rahman N, Vejnar CE, Zdobnov EM (2009) miROrtho: Computational survey of microRNA genes. *Nucleic Acids Res* 37:D111–D117.
38. Klingensmith J, Nusse R (1994) Signaling by wingless in *Drosophila*. *Dev Biol* 166: 396–414.
39. Adams MD, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.
40. Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 100:(Suppl2):14537–14542.
41. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors *in vivo*. *PLoS Biol* 4:e20.
42. Wistrand M, Kall L, Sonnhammer EL (2006) A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci* 15: 509–521.
43. Briscoe AD, Chittka L (2001) The evolution of color vision in insects. *Annu Rev Entomol* 46:471–510.
44. Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB (2007) Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature* 445:86–90.
45. Frommer WB (2010) CO2mmon sense. *Science* 327:275–276.
46. Marella S, et al. (2006) Imaging taste responses in the fly brain reveals a functional map of taste category and behavior. *Neuron* 49:285–295.
47. Chyb S, Dahanukar A, Wickens A, Carlson JR (2003) *Drosophila* Gr5a encodes a taste receptor tuned to trehalose. *Proc Natl Acad Sci USA* 100:(Suppl2):14526–14530.
48. Sanchez-Gracia A, Vieira FG, Rozas J (2009) Molecular evolution of the major chemosensory gene families in insects. *Heredity* 103:208–216.
49. Goberdhan DC, Wilson C (2003) The functions of insulin signaling: Size isn't everything, even in *Drosophila*. *Differentiation* 71:375–397.
50. Oldham S, Hafen E (2003) Insulin/IGF and target of rapamycin signaling: A TOR de force in growth control. *Trends Cell Biol* 13:79–85.
51. 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.
52. Wheeler DE, Buck N, Evans JD (2006) Expression of insulin pathway genes during the period of caste determination in the honey bee, *Apis mellifera*. *Insect Mol Biol* 15: 597–602.
53. Arsic D, Guerin PM (2008) Nutrient content of diet affects the signaling activity of the insulin/target of rapamycin/p70 S6 kinase pathway in the African malaria mosquito *Anopheles gambiae*. *J Insect Physiol* 54:1226–1235.
54. Ranson H, et al. (2002) Evolution of supergene families associated with insecticide resistance. *Science* 298:179–181.
55. Enayati AA, Ranson H, Hemingway J (2005) Insect glutathione transferases and insecticide resistance. *Insect Mol Biol* 14:3–8.
56. Page RD, Lee PL, Becher SA, Griffiths R, Clayton DH (1998) A different tempo of mitochondrial DNA evolution in birds and their parasitic lice. *Mol Phylogenet Evol* 9: 276–293.
57. Blanc G, et al. (2007) Reductive genome evolution from the mother of *Rickettsia*. *PLoS Genet* 3:e14.
58. Andersson SG, et al. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396:133–140.
59. Ogata H, et al. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* 293:2093–2098.
60. Lescot M, et al. (2008) The genome of *Borrelia recurrentis*, the agent of deadly louse-borne relapsing fever, is a degraded subset of tick-borne *Borrelia duttonii*. *PLoS Genet* 4:e1000185.
61. Alsmark CM, et al. (2004) The louse-borne human pathogen *Bartonella quintana* is a genomic derivative of the zoonotic agent *Bartonella henselae*. *Proc Natl Acad Sci USA* 101:9716–9721.
62. Fournier PE, Suhre K, Fournous G, Raoult D (2006) Estimation of prokaryote genomic DNA G+C content by sequencing universally conserved genes. *Int J Syst Evol Microbiol* 56:1025–1029.
63. Fournier PE, et al. (2006) Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLoS Genet* 2:e7.
64. Vallenet D, et al. (2008) Comparative analysis of Acinetobacters: Three genomes for three lifestyles. *PLoS One* 3:e1805.
65. Culpepper GH (1944) The rearing and maintenance of a laboratory colony of the body louse. *Am J Trop Med Hyg* 24:327–329.
66. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5: e254.
67. Myers EW, et al. (2000) A whole-genome assembly of *Drosophila*. *Science* 287: 2196–2204.
68. Venter JC, et al. (2001) The sequence of the human genome. *Science* 291:1304–1351.

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., TTTGGTTTTTTTTTTTTTTGGATTGGTTTTTTTTTT). 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.

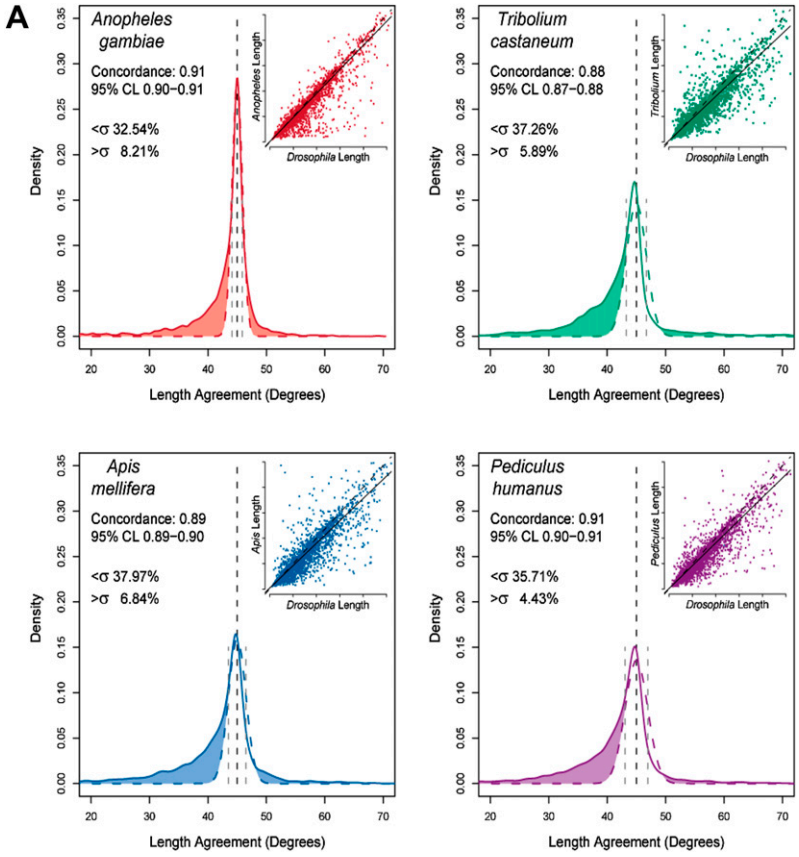

Fig. S1. (Continued)

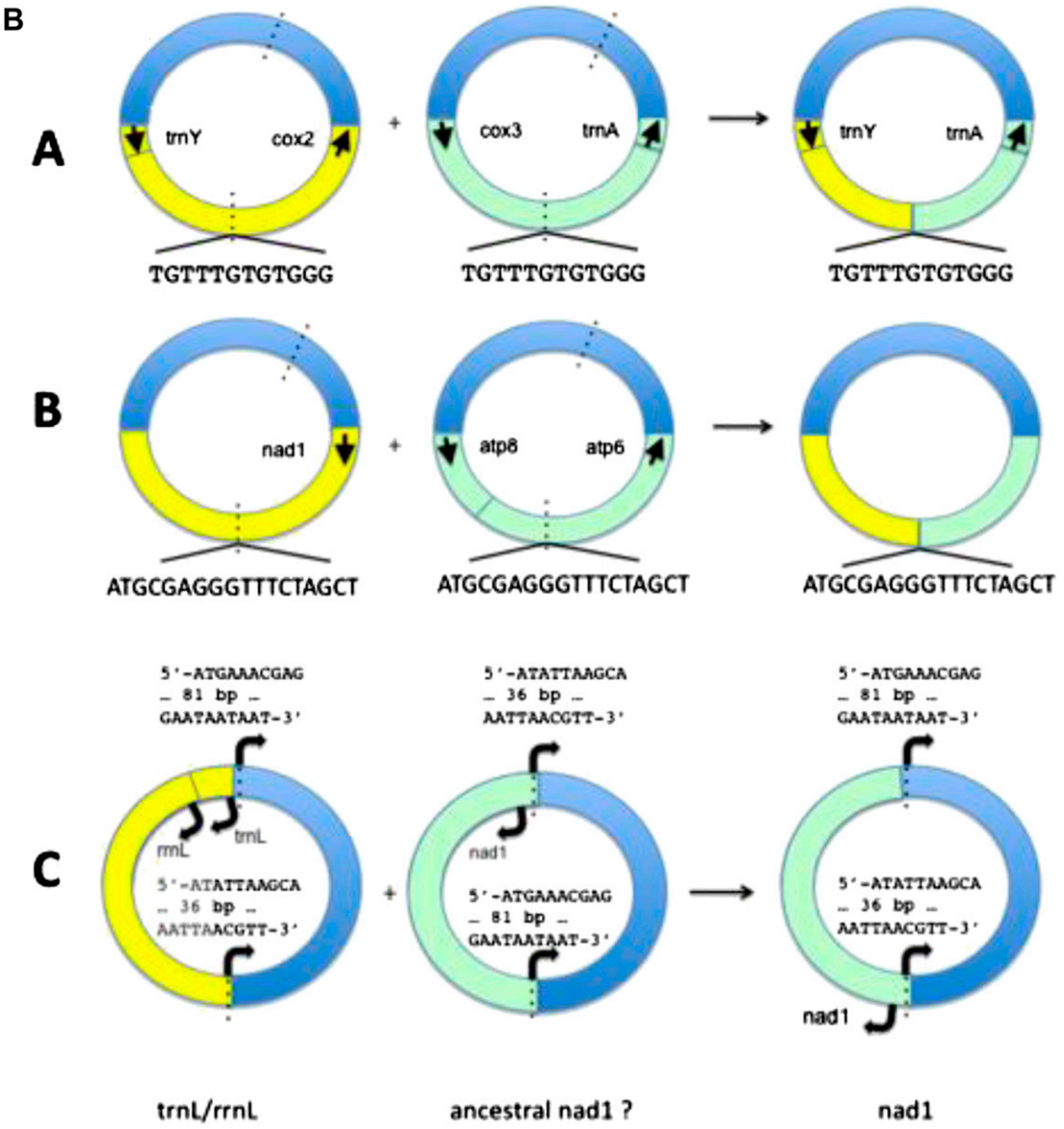

Fig. S1. (*A*) Orthologous protein-length analysis. Orthologous protein-length agreement between *Drosophila melanogaster* proteins with single-copy orthologs in four other insect species: *Anopheles gambiae* (red), *Tribolium castaneum* (green), *Apis mellifera* (blue), and *Pediculus humanus humanus* (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*).

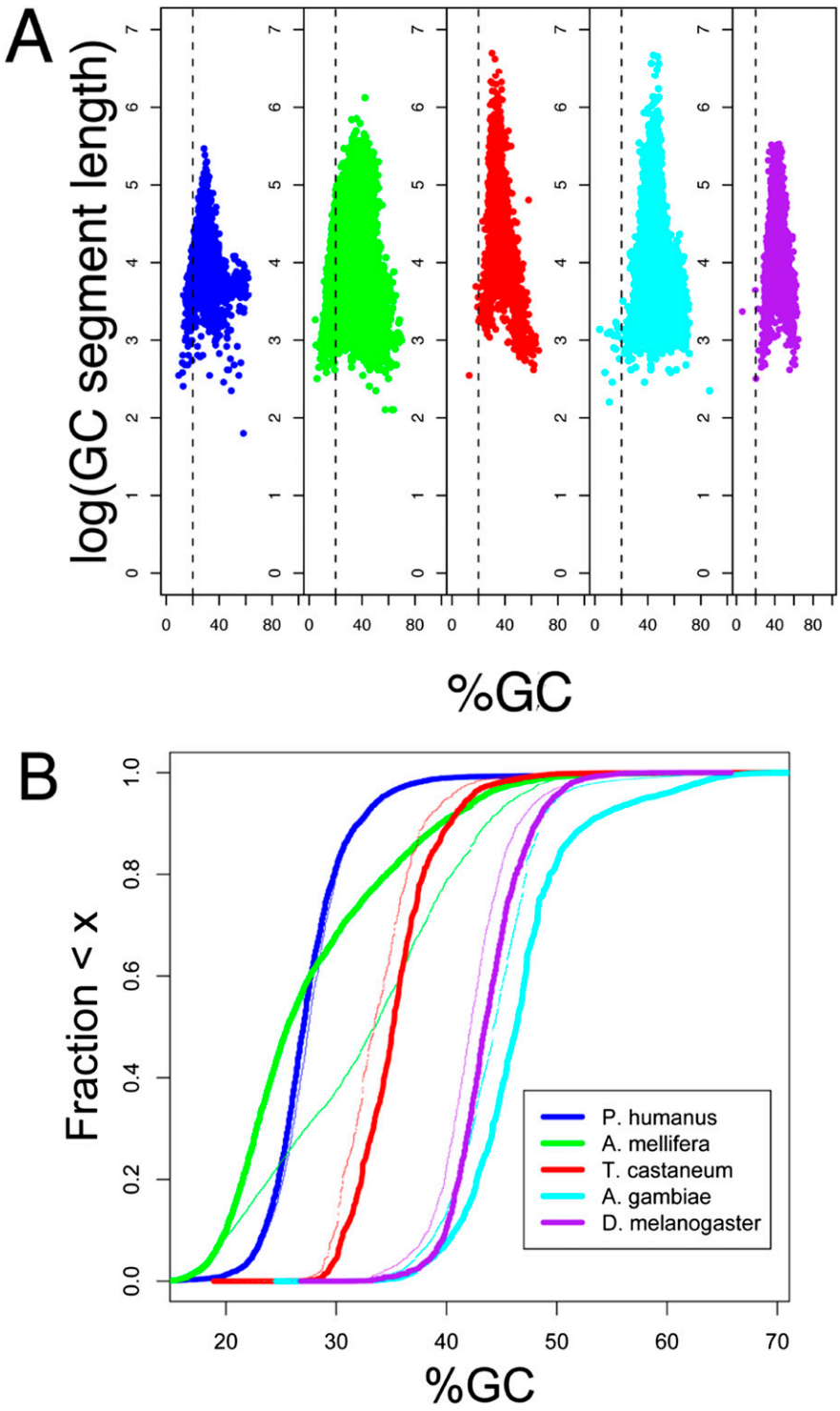

Fig. S2. Comparison of GC-content domains in the insects *Pediculus humanus humanus*, *Apis mellifera*, *Tribolium castaneum*, *Anopheles gambiae*, and *Drosophila melanogaster*. (*A*) GC-content domain lengths versus GC percentage. Hatched line at 20% shown for comparison. (*B*) *P. h. humanus* genes 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%).

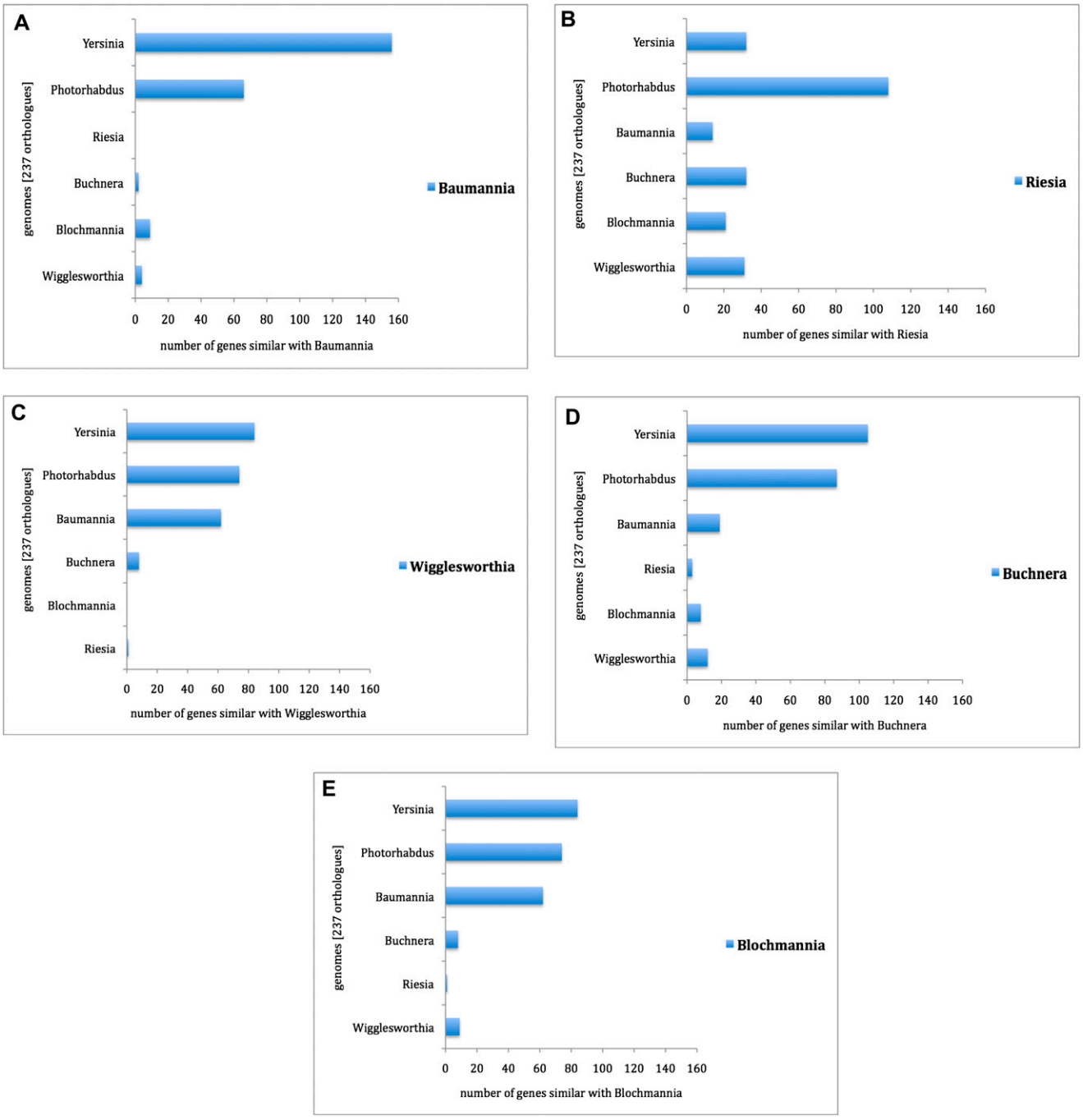

Fig. S3. A genome-wide comparison of *Candidatus* Riesia pediculicola with the primary endosymbionts, *Wigglesworthia glossinida* (tsetse flies), *Blochmannia floridanus* (not shown), *B. pennsylvanicus* (carpenter ants), the automonous *Buchnera aphidicola* (aphids) strains APS and BBp, Sg (not shown), *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.*

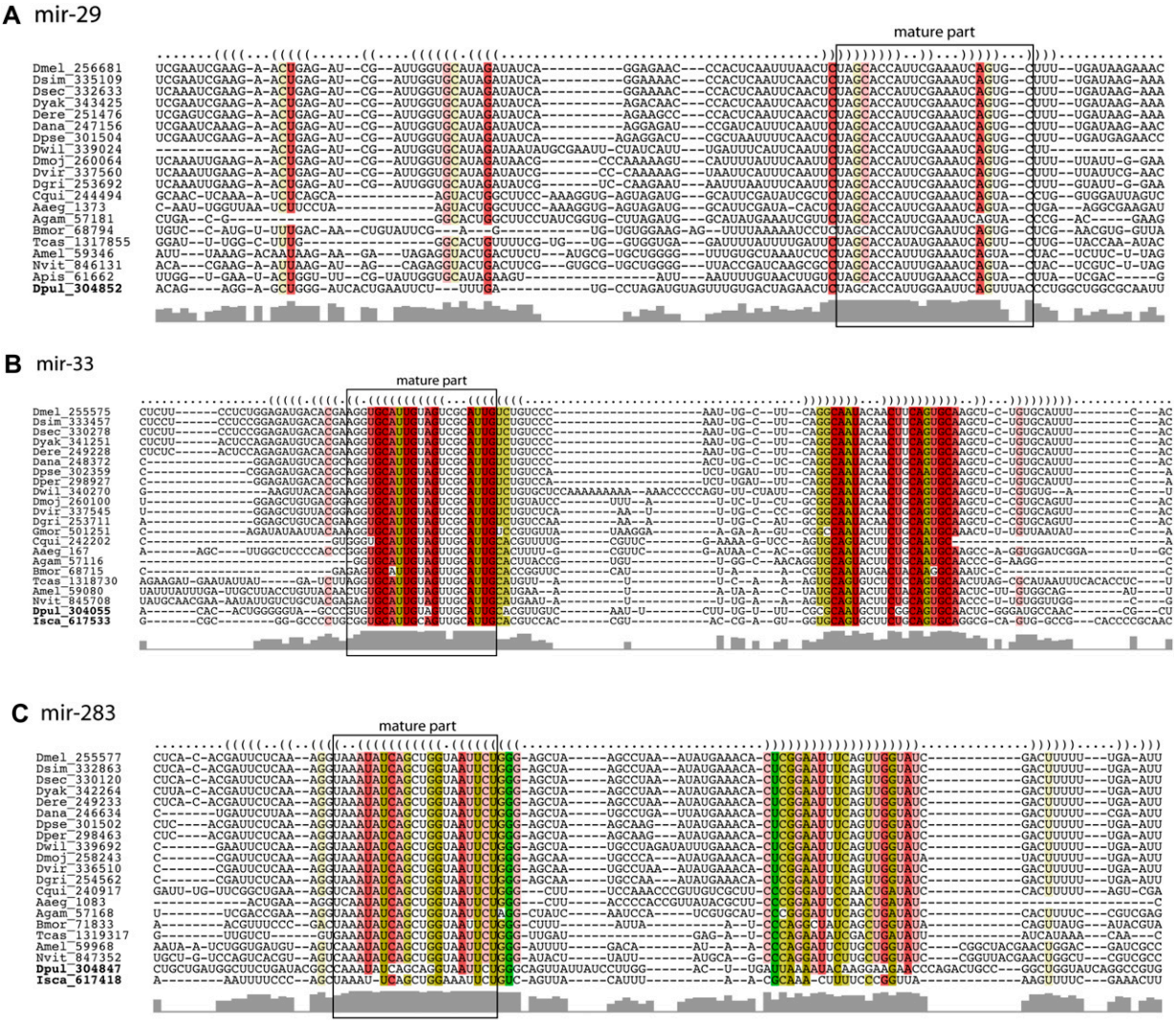

Fig. S4. (Continued)

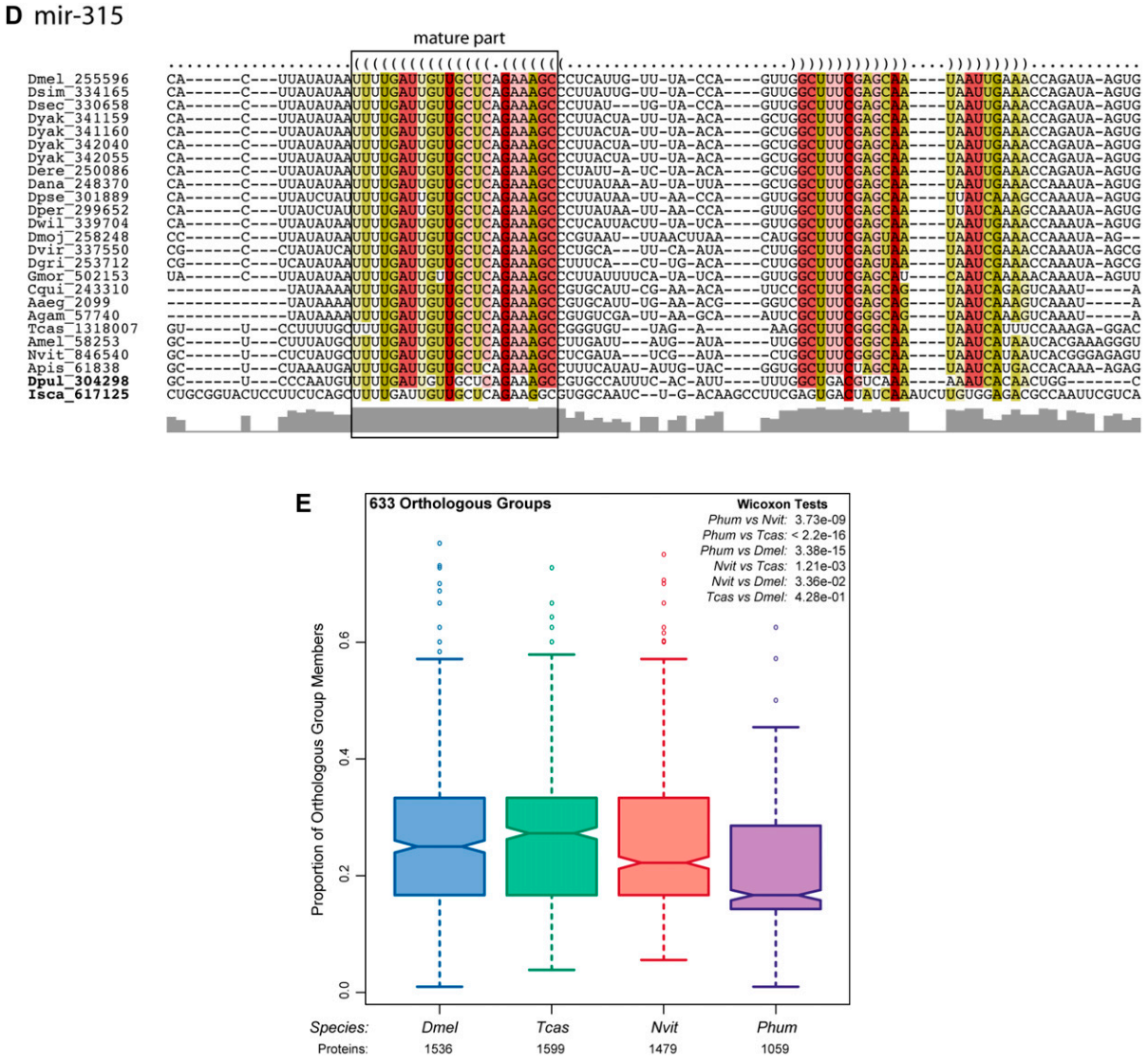

Fig. S4. (Continued)

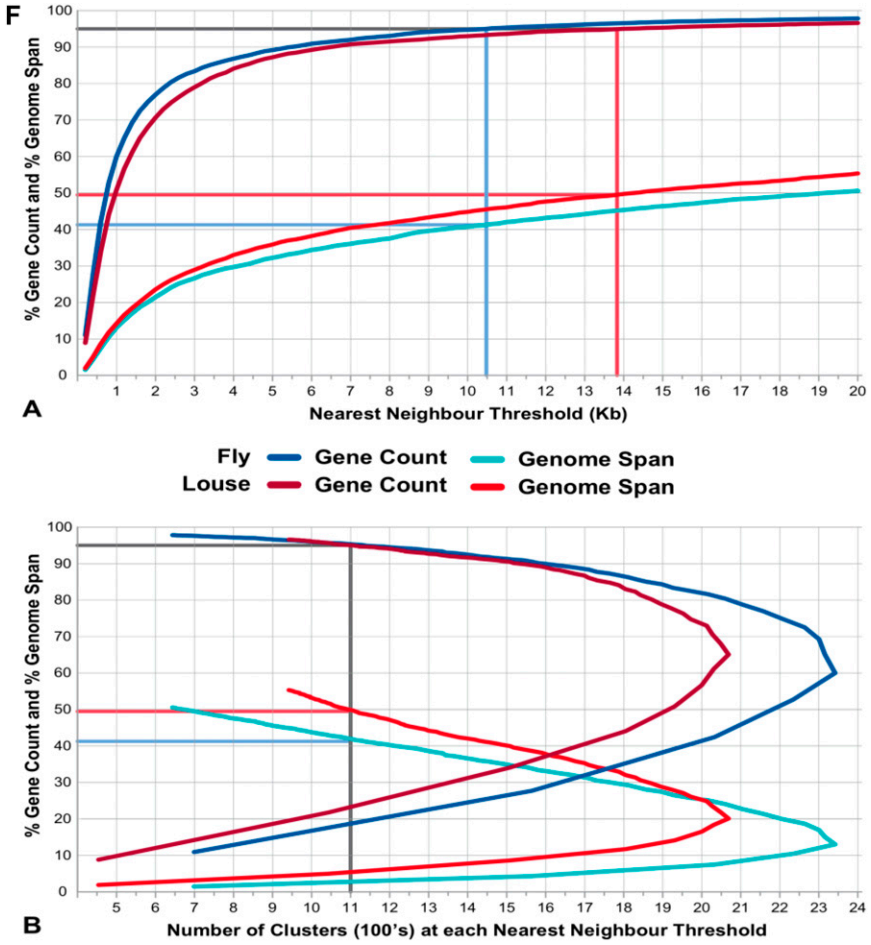

Fig. S4. Multiple alignment of microRNA genes well-represented in insect genomes and found in at least a few more basal lineages (e.g., crustaceans; shown in bold) that we failed to identify in both the *Pediculus humanus humanus* genome and raw sequencing reads; this suggests an evolutionary loss of 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)