4-2015

Comparison of The Genome Profiles Between Head and Body Lice

Jae Soon Kang
Korea Institute of Toxicology

Yong-Ho Cho
Seoul National University

Ju Hyeon Kim
Seoul National University

Sang Hyeon Kim
Seoul National University

Seungil Yoo
Insilicogen, Inc.

See next page for additional authors

Authors
Jae Soon Kang, Yong-Ho Cho, Ju Hyeon Kim, Sang Hyeon Kim, Seungil Yoo, Seung-Jae Noh, Junhyung Park, Kyong-Sup Yoon, J. Marshall Clark, Barry R. Pittendrigh, Jongsik Chun, and Si Hyeock Lee

Cover Page Footnote
This is an accepted manuscript of an article published by Elsevier in the *Journal of Asia-Pacific Entomology*, available online at http://dx.doi.org/10.1016/j.aspen.2015.04.010

Comparison of the genome profiles between head and body lice

Jae Soon Kang, Yong-Jun Cho, Ju Hyeon Kim, Sang Hyeon Kim, Seungil Yoo, Seung-Jae Noh, Junhyung Park, Kyong Sup Yoon, J. Marshall Clark, Barry R. Pittendrigh, Jongsik Chun, Si Hyeock Lee

PII: S1226-8615(15)00050-3
DOI: doi: 10.1016/j.aspen.2015.04.010
Reference: ASPEN 645

To appear in: *Journal of Asia-Pacific Entomology*

Received date: 26 February 2015
Revised date: 14 April 2015
Accepted date: 16 April 2015

Please cite this article as: Kang, Jae Soon, Cho, Yong-Jun, Kim, Ju Hyeon, Kim, Sang Hyeon, Yoo, Seungil, Noh, Seung-Jae, Park, Junhyung, Yoon, Kyong Sup, Marshall Clark, J., Pittendrigh, Barry R., Chun, Jongsik, Lee, Si Hyeock, Comparison of the genome profiles between head and body lice, *Journal of Asia-Pacific Entomology* (2015), doi: 10.1016/j.aspen.2015.04.010

Comparison of the genome profiles between head and body lice

Jae Soon Kang [a], Yong-Jun Cho [b], Ju Hyeon Kim [c], Sang Hyeon Kim [c], Seungil Yoo [d], Seung-Jae Noh [d], Junhyung Park [d], Kyong Sup Yoon [e], J. Marshall Clark [f], Barry R. Pittendrigh [g], Jongsik Chun [b] and Si Hyeock Lee [c,h,*]

[a] Gyeongnam Department of Environmental Toxicology and Chemistry, Korea Institute of Toxicology, Jin-Ju, Gyeongnam, Korea

[b] School of Biological Sciences, Seoul National University, Seoul, Korea

[c] Department of Agricultural Biotechnology, Seoul National University, Seoul, Korea

[d] Department of Research, Codes Division, Insilicogen, Inc., Suwon, 441-813, Korea

[e] Department of Biological Sciences and Environmental Sciences Program, Southern Illinois University-Edwardsville, Edwardsville, IL, 62026, USA

[f] Department of Veterinary & Animal Science, University of Massachusetts, Amherst, MA 01003, USA

[g] Department of Entomology, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA

[h] Research Institute for Agriculture and Life Science, Seoul National University, Seoul, Korea

Running title: Genome profiles of head louse

* Corresponding author: Department of Agricultural Biotechnology, Seoul National University, 151-921, Seoul, Korea. Tel.: +82 2 880 4704; Fax: +82 2 873 2319; E-mail address: shlee22@snu.ac.kr (S.H. Lee)

Abstract

The body louse (*Pediculus humanus humanus*) is known to have diverged from the head louse (*P. humanus capitis*) but genomic differences between these two subspecies still remain unexplored. To compare genomic profiles between head and body lice, whole genome sequences of head lice were determined by next generation sequencing methods based on both Illumina Genome analyzer and Roche GS FLX pyrosequencing and compared with the reference genome sequences of the body louse. Total consensuses generated by mapping to the body louse genome in conjunction with *de novo* assembly of head louse genome sequences revealed a head louse genome size of 110 Mbp with a 96% coverage of the body louse genome sequences. A total of 12,651 genes were predicted from the head louse genome sequences although more precise assembly and functional annotation of the genome is required for a more accurate gene count. Among the 873 genes that were putatively specific to the head louse, 15 genes were confirmed to be transcribed in both head and body lice, suggesting the previously estimated gene number of the body louse was likely underestimated. The single nucleotide polymorphism analysis showed that the nucleotide diversity of genome between head and body lice was 2.2%, which was larger than that of the transcriptome between head and body lice. An endosymbiont genome analysis showed that the composition of endosymbionts in head lice was similar to that of body lice and *Candidatus* Riesia pediculicola was the primary endosymbiont in both head and body lice.

Key words: *Pediculus humanus*, louse, genome sequencing, gene homologue analysis, SNP analysis, endosymbiont

Introduction

Both the head louse (*Pediculus humanus capitis*) and the body louse (*P. humanus humanus*) are obligatory human ectoparasites feeding exclusively human blood. It has been suggested that they adapted to human when the common ancestor of human and chimpanzee diverged 5-6 million years ago (Pennisi, 2004). The body louse is speculated to have diverged from the head louse when human began to wear clothing (Kittler et al., 2003). However, the taxonomic status of the head and body lice is still disputable because fertile F1 hybrid can be generated between head and body lice under laboratory conditions although their interbreeding has not been observed in the wild (Mullen and Durden, 2009). Comparison of several molecular markers, such as mitochondrial DNA, nuclear ribosomal DNA and microsatellite DNA (Kittler et al., 2003; Leo and Barker, 2005; Leo et al., 2002; Leo et al., 2005; Light et al., 2008; Reed et al., 2004), suggested that head and body lice are conspecific except for the study using microsatellite DNA, in which head and body lice were proposed to be separate species (Leo et al., 2005).

Despite a similar genetic background, head and body lice have several differences in their biological features, such as niche, body size and vector competence. Head lice live only on the human scalp throughout its entire lifespan whereas body lice live primarily on clothes as well as on body hair except for hair on the scalp (Kittler et al., 2003). Body lice can transmit pathogenic bacteria to human, such as *Rickettsia prowazekii* (epidemic typhus), *Bartonella quintana* (trench fever), and *Borrelia recurrentis* (relapsing fever) (Brouqui et al., 1999; Raoult and Roux, 1999; Rydkina et al., 1999). In contrast, the head louse is not known to transmit pathogen to humans. Because of such differences, these two species are regarded as

the appropriate models for studies of species differentiation and differential vector competence (Kim et al., 2011).

Recently, whole genome sequencing of the body louse was completed (Kirkness et al., 2010). The body louse has a 108 Mb genome, which is the smallest among insect genomes sequenced, and includes 10,773 protein-coding genes. This reduced number of genes in the body louse has been attributed to its simple life style, which includes feeding only on fresh human blood and having humans as a sole host. Similarly, it was reported that the number of immune-related genes in the body louse was less than that of other insects, such as *Drosophila melanogaster*, *Bombyx mori*, *Anopheles gambiae* and *Tribolium castaneum* (Kim et al., 2011). Since the body louse diverged from the head louse relatively recently, the genetic background of the head louse was assumed to be similar to the body louse, thus likely having almost identical genome size, gene number and genome structure.

The transcriptional profiles between body and head lice were recently compared (Olds et al., 2012). Among the 10,775 protein-coding genes predicted from the body louse genome, almost the same number of genes was annotated both in the head louse (10,770 genes) and the body louse (10,771 genes) transcriptomes. Among the 544 genes in the genome of *Candidatus* Riesia pediculicola, a primary endosymbiont of both head and body lice (Sasaki-Fukatsu et al., 2006), 539 genes were observed from the head louse transcriptome, which were similar to the 538 genes observed in the body louse transcriptome. These results suggested that the phenotypic differences between head and body lice were not likely due to different gene components but rather due to differential gene regulation of similar gene sets. To confirm this assumption, it is necessary to determine the whole genome sequences of the head louse and compare them with those of the body louse.

In this study, the whole genome sequencing of the head louse was performed using two next generation sequencing (NGS) methods, Genome analyzer IIx-platform sequencing and GS FLX Titanium-platform sequencing. The sequences generated from NGS were mapped to the genome of the body louse or *de novo* assembled. From this data set, features of the head louse genome were examined, single nucleotide polymorphisms (SNP) analyzed and putative genes predicted. These predicted genes were homologue-compared to the genes of the body louse, from which the head louse-specific genes were identified. In addition, the bacterial endosymbiont community of the head louse was analyzed and the genome of *Ca.* R. pediculicola in the head louse was sequenced and compared to that determined in the body louse.

Materials and methods

Head and body lice rearing

A highly inbred BR-HL strain of head lice was used for the whole genome sequence analysis. The BR-HL strain was originally collected in Bristol, UK, and has been reared on the *in vitro* rearing system (Yoon et al., 2006). For the reference genomic DNA extraction and cDNA preparation, another head louse strain (CA-HL, originally collected from Cambodia) and two body louse strains (SF-BL, collected from San Francisco; CP-BL, Culpepper strain that was used for the body louse genome analysis) were also used. The louse colonies were maintained under conditions of 30°C, 70-80% RH and 16L: 8D in a rearing chamber.

Genomic DNA extraction

Genomic DNA was extracted from approximately 500 newly hatched first instar nymphs before their first blood meal using DNeasy blood & tissue kit (Qiagen, Hilden, Germany). Both quality and quantity of genomic DNA were analyzed by gel electrophoresis and Quant-iT™ PicoGreen® dsDNA Quanititation reagent (Invitrogen, Carlsbad, CA, USA).

Whole genome sequencing

Genome Analyzer IIx (Illumina, San Diego, CA, USA) with a mean length of 101 bp paired-end and GS FLX Titanium (Roche, Indianapolis, CA, USA) using a 3 kb library were used for genome sequencing according to the manufacturer's recommendations at the National Instrumentation Center for Environmental Management (NICEM, Seoul, Korea) and DNA Link, Inc. (Seoul, Korea), respectively. The resulting sequence data were mapped to

8,588 contigs of the body louse genome. Unmapped reads were used for creating *de novo* assembly. All analyses and statistics for *de novo* assembly and reference mapping were performed using CLC Genomics Workbench (CLC bio, Aarhus, Denmark). The gene coding regions were predicted by GeneMark-ES version 2.3a (Ter-Hovhannisyan et al., 2008) in the *de novo* assembled contigs. Predicted protein-encoding genes were analyzed by BLASTP and the ones, which showed significant BLAST similarity (e-value < 10^{-5}) to proteins from other organisms in the non-redundant (NR) database at the National Center for Biotechnology Information (NCBI), were annotated. Comparison of the genes between head and body lice were conducted by BLAST-searching of all genes of each species against opposite species genome database on the condition of < 10^{-4} e-value.

Endosymbiont genomes were annotated using RAST server (Aziz et al., 2008). The reads that unmapped to the genome of the body louse were assembled and the assembled contigs containing 16S ribosomal RNA (16S rRNA) fragment were selected. Selected contigs were used for analyzing the bacterial community by CLcommunity™ (Ver. 2.04, CLC bio). In addition, the genome of *Ca.* R. pediculicola in the head louse was sequenced and mapped to the genome of *Ca.* R. pediculicola in the body louse.

Verification of newly identified putative genes

Among the putative head louse-specific genes, 30 genes, which were not identified from the body louse genome, were further analyzed by PCR to confirm their presence in the genomes of head and body lice. Primer pairs were designed from the putative exon regions of each gene (Supplementary Table 1) and used for PCR using either genomic DNA or cDNA . Genomic DNA was extracted from the BR-HL, CA-HL, SF-BL, and CP-BL. cDNA was

synthesized from the total RNA extracted from the same strains of head and body lice. PCR was conducted with Ex Taq polymerase (Takara Korea Biomedical Inc., Seoul, Korea) and 5 pmole primers under the following thermal program: an initial denaturation at 95°C for 2 min and a total of 34 cycles of 95°C for 20 sec, 52°C for 10 sec, and 72°C for 1 min.

Calculation of Ka/Ks ratios

Through reciprocal blast-searching of protein sequences between head and body lice, 9,015 pairs of orthologous gene sequences were extracted. Each of these pairs was aligned using ClustalW2. The ratios (*Ka/Ks*) of the non-synonymous substitutions per site (*Ka*) to synonymous substitutions per site (*Ks*) were estimated for each orthologue pair and averaged over the entire alignment using Ka/Ks Calculator v.2 (Zhang et al, 2006). Except for 2,303 pairs that failed because of large gap opening in the sequence alignment, 6,721 orthologue pairs were successfully analyzed using the Ka/Ks Calculator. The Ka/Ks Calculator adopts different models for codon substitutions, such as approximate methods (NG, LPB, MYN, etc.) and maximum likelihood methods (GY, MS, MA, etc.), among which the NG method was used to estimate *Ka/Ks* ratios.

Results

Head louse genome features

A total of 11.5 Gb nucleotide sequences (ca. 114 million reads having 101 bp) were obtained from Illumina GA platform (Table 1). These reads, which approximated a 50× average coverage of head louse genome, were mapped to the body louse genome composed of 8,588 contigs. A total of 91.5 million reads were mapped to the body louse genome and 22.5 million reads were unmapped. The unmapped reads were undergone *de novo* assembly, from which 7.6 million reads were assembled. From reference mapping and *de novo* assembly, 10,621 contigs (8,555 reference mapping contigs and 2,066 *de novo* assembled contigs) were generated. A total of 250 Mb nucleotide sequences (the average size of a read was approximately 340 bp) were additionally obtained from Roche GS FLX platform to compensate for gaps between contigs generated by sequences mapping and *de novo* assembly. Through assembly of contigs assembled from Illumina raw data and reads generated from Roche GS FLX, 1,375 supercontigs were finally obtained. The fraction of the reference body louse genome covered by head louse sequences was 96% and the head louse genome size was determined to be approximately 110 Mb.

The estimated features of the head louse genome were compared with those of the body louse as summarized in Table 2. Predictions using the aforementioned computational programs determined 12,651 protein-coding genes, which were far more than the 10,775 body louse genes initially predicted. A total of 169 transfer ribonucleic acids (tRNAs) were determined from the head louse, which were comparable with 161 tRNA genes from the body louse. Only 21 microRNAs (miRNAs) were found in the head louse compared with 57

miRNA genes in the body louse. The head louse had 180,785 tandem repeats, which were more than that determined in body lice (130,608). The average GC content of the head louse was 26.9%, which was similar to that of the body louse (28%). Most genome features predicted in the head louse were in excess compared with those determined in the body louse, whereas the number of miRNA determined in the head louse was less than that determined in the body louse.

Gene homologue analysis

Homologue analyses were performed so that each gene sequence in the head louse was matched to the corresponding gene sequence in the body louse by using BLASTP. Of 12,651 head louse genes, 873 genes were determined to be specific to the head louse and 11,778 genes were homologous to the body louse. In contrast, 422 genes were specific to the body louse among the 10,773 body louse genes and 10,351 genes were homologous to head louse genes. Among the putative head louse-specific genes, 61 genes were identified by BLASTP search but 812 genes were only identified as hypothetical proteins. As most genes had high e-values ($> 10^{-4}$), it is likely that they may have been generated by prediction error. The 61 putative head louse-specific genes were manually re-analyzed through BLASTP search, among which 30 genes were not found in the body louse genome. Among these 30 genes, only 15 genes exhibited good agreements between their predicted exon-intron structure and observed transcript. All these 15 genes, however, were also detected in body louse cDNA as judged by PCR, which demonstrates that they are present in both head and body lice (Table 4). This finding further suggests that the previously estimated gene number of the body louse was likely underestimated.

SNP analysis

Based on the mapping results of the genome sequences generated by Illumina GA platform, we analyzed nucleotide variations (insertions, deletions and SNPs) of the head louse genome sequence against the body louse genome sequence (Table 3). The total amounts of insertions, deletions and SNPs observed in the head louse genome were approximately 0.27 Mb, 0.68 Mb and 1.45 Mb, respectively, all of which was equivalent to 2.2% of total genome. Among the 8,555 contigs, the highest nucleotide diversity was 8.13% and 4,603 contigs (equivalent to 59.5% of whole genome size) showed ≥ 2.0% nucleotide diversity. Only 40 contigs from the head louse were identical to those of the body louse. On average, 280 bp nucleotide variations were generated in a contig, in which SNPs occupied approximately 170 bp.

The majority (90.9%) of SNPs in the head louse genome were generated in non-coding regions whereas the remaining 9.1% were found in coding regions (i.e. cSNPs). Among the cSNPs observed, 91.5% were synonymous whereas 8.5% were non-synonymous (4.6 % amino-acid substitution, 0.2 % termination and 3.7 % frame-shift). To assess the adaptive evolution between head and body lice, we analyzed the *Ka/Ks* ratios of 6,721 orthologue pairs. The majority of these genes (4,027, 60%) had *Ka/Ks* ratios of ≤ 0.3 ($p \leq 0.05$), indicating strong purifying selection and conservation of protein structure. About 35% of the genes had *Ka/Ks* ratios of 0.3~1.0, indicating neutral selection. Five genes had *Ka/Ks* ratio values significantly greater than 1, indicative of positive selection. These five genes were identified to code for (1) hypothetical protein (PHUM497030, *Ka/Ks*=9.723), (2) putative Rho-GTPase-activating protein (PHUM131520, *Ka/Ks*=3.884), (3) putative voltage-gated

potassium channel (PHUM065240, *Ka/Ks*=3.616), (4) putative arginine/serine-rich protein (PHUM106870, *Ka/Ks*=2.821), and (5) predicted protein (PHUM131530, *Ka/Ks*=2.312).

Immune related genes

All the immune related genes previously identified in the body louse genome (Kim et al., 2011) were also annotated in the head louse genome (Supplementary Table 2). The average nucleotide sequence identity in 1:1 orthologous genes between head and body lice was 96.8%.

Endosymbiont analysis

The primary bacterial endosymbiont species in head lice was determined to be *Ca.* R. pediculicola (67.8%), which was the major endosymbiont in body lice, followed by another *Ca.* Riesia sp. (10.5%) and unidentified species belonging to family Enterobacteriaceae (5.4%) (Fig. 1). The reads that were not mapped to the body louse genome were assembled, from which six contigs were generated. From these contigs, a total of 520 genes were found, of which 513 genes were found in the *Ca.* R. pediculicola genome of the body louse. Therefore, 7 genes identified in the *Ca*. R. pediculicola genome from head lice were not found in the *Ca*. R. pediculicola genome from body lice. The homology of nucleotide sequence between *Ca.* R. pediculicola of head and body lice was 99.64%.

Discussion

Only 96% of the head louse genome sequences could be mapped to the body louse genome. This incomplete mapping appears to be due, in part, to the sub-optimized reference body louse genome that still has gaps (Kirkness et al., 2011) and to the inaccurate assembly of highly variable sequences obtained from the head louse genomic DNA, which appears to be heterozygous in nature. A total of 12,651 genes were predicted from the head louse genome, which was unexpectedly higher than the gene number predicted from the body louse genome (10,775 genes). Considering the extremely high transcriptome similarity between head and body lice (Olds et al., 2012), such a large increase in the number of genes in the head louse genome appears substantially artifactual and likely due to errors in sequence assembly and annotation of short reads, particularly those generated from heterozygous genome sequences. Among the 873 putative head louse-specific genes, most were assumed to be generated by prediction error as they showed high e-values ($> 10^{-2}$) to hypothetical proteins. Nevertheless, 15 of the genes were confirmed to be transcribed in both head and body lice, suggesting that they are actually present in both genomes and the previously estimated gene number of the body louse genome is likely underestimated. The recent re-analysis of the honey bee genome revealed ~5000 more protein-coding genes compared with the previously reported gene set, where the finding of such additional genes was due to improved assembly and updated evidenced gene data including new RNAseq and protein data (Elsik et al., 2014). If this is the case, the genomes of head and body lice possess at least 10,790 genes and the actual gene number would be expected to increase with improved assembly and annotation.

In homologue analysis of genes between head and body lice, 422 body louse-specific genes were annotated. Considering that no body louse-specific transcript was observed in the comparison of transcriptomes between head and body lice (Olds et al., 2012), it is highly likely that misidentification of body louse-specific genes were due to the incomplete assembly of head louse genome and the sub-optimized gene prediction from the head louse genome.

Vector competence is one of most notable differences between head and body lice. The differences in the immune response between head and body lice were regarded as primary factors determining their differential vector competence (Kim et al., 2011). A total of 93 immune-related genes were annotated from the body louse genome (Kim et al., 2011) and the existence of these genes was confirmed in head lice by transcriptional profiling study (Olds et al., 2012). Homologue analysis of all immune related genes in the genomes between head and body lice confirmed that both head and body lice shared 1:1 orthologous genes and that IMD, GNBP and FADD were not present either head or body lice. Nevertheless, the relative transcriptional level analysis using quantitative real-time PCR showed that some immune-related genes, such as peptidoglycan recognition protein (PGRP) and defensin-1, had higher basal transcription levels in head lice than in body lice (Kim et al., 2012). Since no copy number differences of these genes were observed between head and body lice, their differential transcription is primarily attributable to different gene regulation factors in non-coding region, such as *cis*-/*trans*-regulatory elements and miRNAs. In addition, low nucleotide diversity was observed in the coding regions, whereas higher nucleotide diversity was mainly exhibited in the non-coding regions. These results further suggest that the differences between head and body lice are likely generated by gene regulation rather than

gene composition although the possibility of functional alteration of some genes by cSNPs cannot be completely ruled out. Taken together, in-depth comparison of regulation factors in non-coding region, including miRNA, and investigation on the cSNPs causing the functional alteration in encoded proteins would contribute to understanding the evolutionary divergence that explains the biological and physiological differences between head and body lice.

The transcriptional comparison between head and body lice showed that the nucleotide diversity of the two transcriptomes was also low (0.1-1.3%) (Olds et al., 2012). However, when collectively comparing the whole genomes, including both coding and non-coding regions, the nucleotide diversity between head and body lice increased to 0-8.13% (average 2.2%). In a previous study investigating the sequence diversities in the mitochondrial cytochrome c oxidase I and II markers within each of 23 insect species and between closely related species, the intraspecific nucleotide diversities were in the range of 0.03-2.71% whereas the interspecific nucleotide diversities were in the range of 0.18-6.76% (Roe and Sperling, 2007). Another study using nuclear DNA markers of three *Drosophila* species, including coding and non-coding regions, also reported that their intraspecific nucleotide diversity were 0.4-2.0% depending on species and region (Moriyama and Powell, 1996). Taken together, the nucleotide diversity level between head and body lice is in between the intraspecific and interspecific boundaries, further suggesting that head and body lice are evolving to separate species from their status of con-species (Leo et al., 2002) or ecotypes of the same species (Li et al., 2010).

Among the SNPs found in the coding regions, high *Ka/Ks* values (>1) were observed in five genes, including a hypothetical protein (PHUM497030), a putative Rho-GTPase-activating protein (PHUM131520), a putative voltage-gated potassium channel

(PHUM065240), a putative arginine/serine-rich protein (PHUM106870) and a hypothetical protein (PHUM131530). This finding suggests these genes have been under strong positive selection during the evolution of body lice from head lice. It would be interesting to study the relationships between the functional divergence of these five genes and the adaptive evolution of head and body lice.

In many insects, endosymbionts provide various important functions necessary for the survival of their host. The best known insect endosymbiont is *Buchnera* in the pea aphid (Shigenobu et al., 2000). Both head and body lice also have an obligate bacterial endosymbiont, *Ca.* R. pediculicola, which carries out the critical function of the biosynthesis of pantothenic acid (Perotti et al., 2009). Because the louse host and endosymbiont have coevolved over a long period of time, a comparison of the endosymbiont community and genome of their primary endosymbiont may be necessary for understanding the evolutionary processes leading to several differences apparent between head and body lice. In the endosymbiont community analysis, *Ca.* R. pediculicola was found to be the primary endosymbiont occupying 67.8% of head louse endosymbiont community with another *Ca.* Riesia species occupying an additional 10.5%. Although there is little information on the minor species of these endosymbionts, they may affect some of the the physiological differences seen between head and body lice if they differ in their composition. Thus, further analysis of the endosymbiont community in both head and body lice may provide crucial information to understand the biological differences between these lice.

A total of 520 genes were predicted in the *Ca.* R. pediculicola genome from head lice, of which 513 genes were also found in the *C. Riesia pediculicola* genome of body louse. Based on these 513 orthologous genes, the nucleotide divergence of *Ca.* R. pediculicola was 0.36%

between head and body lice. This result was similar to the finding of a previous divergence study using 16S rRNA, where each of the major endosymbionts from head and body lice showed 0.31-0.33% nucleotide diversity (Allen et al., 2007). This level of nucleotide diversity is lower than that associated with either the genome or transcriptome analysis between head and body lice, and indicates that *Ca.* R. pediculicola was acquired before head and body lice diverged.

Acknowledgements

This work was supported by a grant from the NIH/NIAID (5 R01 AI045062-06) to JMC and SHL. JH Kim and SH Kim were supported in part by Brain Korea 21 program.

References

Allen, J.M., Reed, D.L., Perotti, M.A., Braig, H.R., 2007. Evolutionary relationships of "*Candidatus* Riesia spp.," endosymbiotic *Enterobacteriaceae* living within hematophagous primate lice. Appl Environ Microbiol 73, 1659-1664.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75.

Brouqui, P., Lascola, B., Roux, V., Raoult, D., 1999. Chronic Bartonella quintana bacteremia in homeless patients. N Engl J Med 340, 184-189.

Elsik, C.G., Worley, K.C., Bennett, A.K., Beye, M., Camara, F., Childers, C.P., de Graaf, D.C., Debyser, G., Deng, J.X., Devreese, B., Elhaik, E., Evans, J.D., Foster, L.J., Graur, D., Guigo, R., Hoff, K.J., Holder, M.E., Hudson, M.E., Hunt, G.J., Jiang, H.Y., Joshi, V., Khetani, R.S., Kosarev, P., Kovar, C.L., Ma, J., Maleszka, R., Moritz, R.F.A., Munoz-Torres, M.C., Murphy, T.D., Muzny, D.M., Newsham, I.F., Reese, J.T., Robertson, H.M., Robinson, G.E., Rueppell, O., Solovyev, V., Stanke, M., Stolle, E., Tsuruda, J.M., Van Vaerenbergh, M., Waterhouse, R.M., Weaver, D.B., Whitfield, C.W., Wu, Y.Q., Zdobnov, E.M., Zhang, L., Zhu, D.H., Gibbs, R.A., Teams, H.P., Consor, H.B.G.S., 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. Bmc Genomics 15.

Kim, J.H., Min, J.S., Kang, J.S., Kwon, D.H., Yoon, K.S., Strycharz, J., Koh, Y.H.,

Pittendrigh, B.R., Clark, J.M., Lee, S.H., 2011. Comparison of the humoral and cellular immune responses between body and head lice following bacterial challenge. Insect Biochem Molec 41, 332-339.

Kim, J.H., Yoon, K.S., Previte, D.J., Pittendrigh, B.R., Clark, J.M., Lee, S.H., 2012. Comparison of the immune response in alimentary tract tissues from body versus head lice following *Escherichia coli* oral infection. J Asia Pacific Entomol 15, 409-412.

Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., Gerlach, D., Kriventseva, E.V., Elsik, C.G., Graur, D., Hill, C.A., Veenstra, J.A., Walenz, B., Tubio, J.M., Ribeiro, J.M., Rozas, J., Johnston, J.S., Reese, J.T., Popadic, A., Tojo, M., Raoult, D., Reed, D.L., Tomoyasu, Y., Kraus, E., Mittapalli, O., Margam, V.M., Li, H.M., Meyer, J.M., Johnson, R.M., Romero-Severson, J., Vanzee, J.P., Alvarez-Ponce, D., Vieira, F.G., Aguade, M., Guirao-Rico, S., Anzola, J.M., Yoon, K.S., Strycharz, J.P., Unger, M.F., Christley, S., Lobo, N.F., Seufferheld, M.J., Wang, N., Dasch, G.A., Struchiner, C.J., Madey, G., Hannick, L.I., Bidwell, S., Joardar, V., Caler, E., Shao, R., Barker, S.C., Cameron, S., Bruggner, R.V., Regier, A., Johnson, J., Viswanathan, L., Utterback, T.R., Sutton, G.G., Lawson, D., Waterhouse, R.M., Venter, J.C., Strausberg, R.L., Berenbaum, M.R., Collins, F.H., Zdobnov, E.M., Pittendrigh, B.R., 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A 107, 12168-12173.

Kirkness, E.F., Haas, B.J., Sun, W.L., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., Gerlach, D., Kriventseva, E.V., Elsik, C.G., Graur, D., Hill, C.A., Veenstra, J.A., Walenz, B., Tubio, J.M.C., Ribeiro, J.M.C., Rozas, J.,

Johnston, J.S., Reese, J.T., Popadic, A., Tojo, M., Raoult, D., Reed, D.L., Tomoyasu, Y., Kraus, E., Mittapalli, O., Margam, V.M., Li, H.M., Meyer, J.M., Johnson, R.M., Romero-Severson, J., VanZee, J.P., Alvarez-Ponce, D., Vieira, F.G., Aguade, M., Guirao-Rico, S., Anzola, J.M., Yoon, K.S., Strycharz, J.P., Unger, M.F., Christley, S., Lobo, N.F., Seufferheld, M.J., Wang, N.K., Dasch, G.A., Struchiner, C.J., Madey, G., Hannick, L.I., Bidwell, S., Joardar, V., Caler, E., Shao, R.F., Barker, S.C., Cameron, S., Bruggner, R.V., Regier, A., Johnson, J., Viswanathan, L., Utterback, T.R., Sutton, G.G., Lawson, D., Waterhouse, R.M., Venter, J.C., Strausberg, R.L., Berenbaum, M.R., Collins, F.H., Zdobnov, E.M., Pittendrigh, B.R., 2011. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle (vol 107, pg 12168, 2010). P Natl Acad Sci USA 108, 6335-6336.

Kittler, R., Kayser, M., Stoneking, M., 2003. Molecular evolution of *Pediculus humanus* and the origin of clothing. Curr Biol 13, 1414-1417.

Leo, N.P., Barker, S.C., 2005. Unravelling the evolution of the head lice and body lice of humans. Parasitol Res 98, 44-47.

Leo, N.P., Campbell, N.J., Yang, X., Mumcuoglu, K., Barker, S.C., 2002. Evidence from mitochondrial DNA that head lice and body lice of humans (Phthiraptera: Pediculidae) are conspecific. J Med Entomol 39, 662-666.

Leo, N.P., Hughes, J.M., Yang, X., Poudel, S.K., Brogdon, W.G., Barker, S.C., 2005. The head and body lice of humans are genetically distinct (Insecta: Phthiraptera, Pediculidae): evidence from double infestations. Heredity (Edinb) 95, 34-40.

Li, W., Ortiz, G., Fournier, P.E., Gimenez, G., Reed, D.L., Pittendrigh, B., Raoult, D., 2010. Genotyping of human lice suggests multiple emergencies of body lice from local head louse populations. PLoS Negl Trop Dis 4, e641.

Light, J.E., Toups, M.A., Reed, D.L., 2008. What's in a name: the taxonomic status of human head and body lice. Mol Phylogenet Evol 47, 1203-1216.

Moriyama, E.N., Powell, J.R., 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol Biol Evol 13, 261-277.

Mullen, G., Durden, L.A., 2009. Medical and veterinary entomology, 2nd ed. Academic Press, San Diego.

Olds, B.P., Coates, B.S., Steele, L.D., Sun, W., Agunbiade, T.A., Yoon, K.S., Strycharz, J.P., Lee, S.H., Paige, K.N., Clark, J.M., Pittendrigh, B.R., 2012. Comparison of the transcriptional profiles of head and body lice. Insect Mol Biol 21, 257-268.

Pennisi, E., 2004. Human origins. Louse DNA suggests close contact between early humans. Science 306, 210.

Perotti, M.A., Kirkness, E.F., Braig, H.R., Reed, D.L., 2009. Insect Symbiosis. CRC Press, Boca Raton, FL.

Raoult, D., Roux, V., 1999. The body louse as a vector of reemerging human diseases. Clin Infect Dis 29, 888-911.

Reed, D.L., Smith, V.S., Hammond, S.L., Rogers, A.R., Clayton, D.H., 2004. Genetic analysis of lice supports direct contact between modern and archaic humans. PLoS Biol 2,

e340.

Roe, A.D., Sperling, F.A., 2007. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. Mol Phylogenet Evol 44, 325-345.

Rydkina, E.B., Roux, V., Gagua, E.M., Predtechenski, A.B., Tarasevich, I.V., Raoult, D., 1999. *Bartonella quintana* in body lice collected from homeless persons in Russia. Emerg Infect Dis 5, 176-178.

Sasaki-Fukatsu, K., Koga, R., Nikoh, N., Yoshizawa, K., Kasai, S., Mihara, M., Kobayashi, M., Tomita, T., Fukatsu, T., 2006. Symbiotic bacteria associated with stomach discs of human lice. Appl Environ Microbiol 72, 7349-7352.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H., 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407, 81-86.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., Borodovsky, M., 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18, 1979-1990.

Yoon, K.S., Strycharz, J.P., Gao, J.R., Takano-Lee, M., Edman, J.D., Clark, J.M., 2006. An improved in vitro rearing system for the human louse allows the determination of resistance to formulated pediculicides. Pestic Biochem Physiol 86, 195-202.

Figure captions

Fig. 1. The endsymbiont community of the head louse analyzed by contigs of 16S ribosomal DNA sequences.

Table 1
Summary of reads generated from two NGS methods (Illumina GAIIx platform and Roche GS FLX Titanium platform) and contigs obtained by reference mapping and *de novo* assembly.

Illumina GAIIx	
Total short reads	114,017,738
Reference mapped reads	92,029,559
de novo assembled reads	7,644,178
Unmapped reads	14,978,019
Reference contigs	8,555
de novo contigs	2,066
N50 (total contigs)	33,240
Roche GS FLX Titanium	
Total reads	634,018
Average size of a read (bp)	
Assembled contigs	1,375
Reference coverage	
Total reference length (bp)	110,768,579
GC contents (%)	26.87
Total consensus length (bp)	105,935,226
Fraction of reference covered (%)	96

Table 2
The comparison of genome features between the head louse and the body louse.

Genome feature	Count	Nucleotides (Mb)	Genome fraction (%)
Head louse (Body louse)		114 (110)	100
Protein-coding genes			
Total	12,651 (10,773)	41.5 (33.8)	36.4 (31)
Coding exons	77,709 (69,261)	19.1 (16.6)	16.8 (15)
Introns	65,058 (58,552)	22.4 (17.2)	19.6 (15)
Non-protein-coding genes			
tRNAs	169 (161)	0.012 (0.012)	< 1
miRNAs	21 (57)	0.002 (0.005)	< 1
Tandem repeats	180,785 (130,608)	9 (6.9)	7.9 (6)

* This table form was based on Table 1 of the article by Kirkness et al. (2010).

Table 3
The nucleotide variation of head louse genome sequences mapped to body genome sequences.

	Insertion	Deletion	SNP	Total
Total	270,744	675,356	1,448,591	2,394,691
Average per contig	31.7	79.0	169.3	280.0

The unit of all data in this table is a base pair (bp).

Table 4
The list of additional genes identified in the genomes of both head and body lice.

Gene ID	Gene description	Presence confirmed by PCR	
		Genomic DNA	cDNA
AAZO01000648_ORF00947	Rugose, isoform F	Yes	Yes
AAZO01001842_ORF02612	U7 snRNA-associated Sm-like protein LSm11-like	Yes	Yes
AAZO01002015_ORF02859	Immunoglobulin mu binding protein 2	Yes	Yes
AAZO01002204_ORF03066	General transcription factor IIH subunit 3-like	Yes	Yes
AAZO01002574_ORF03683	Pangolin	Yes	Yes
AAZO01003670_ORF05393	Armadillo repeat-containing protein 2	Yes	Yes
AAZO01003971_ORF05808	Homeobox protein abdominal-A	Yes	Yes
AAZO01005358_ORF01553	Pro-corazonin preproprotein	Yes	Yes
AAZO01005807_ORF02528	Similar to visgun CG16707-PC	Yes	Yes
AAZO01006056_ORF02941	RNA binding protein Vera (Insulin-like growth factor 2)	Yes	Yes
AAZO01006114_ORF03027	RNA-binding protein 38-like isoform 1	Yes	Yes
AAZO01006277_ORF03445	Leucine-rich repeat and WD repeat-containing protein 1	Yes	Yes
AAZO01006574_ORF04009	UPF0459 protein CG10681-like	No	Yes
AAZO01006644_ORF04092	Hypoxia-inducible factor 1 alpha	No	Yes
AAZO01006644_ORF04093	Hypoxia-inducible factor 1 alpha	Yes	Yes

Supplementary Table 1

Sequence information of the primers used for the verification of newly annotated head and body lice genes

Gene ID	Gene description	Primer sequence (5'-3')	Amp size [a]
ORF00947	Rugose, isoform F	F: TTGGTCGGTGGAGAATTCGA R: GATTTGGAGCGCAATGATCC	100
ORF02612	U7 snRNA-associated Sm-like protein LSm11-like	F: ACGAGAGGTCCGAGAGAAAT R: GCCAAACTTCTTGCACGTCT	100
ORF02859	Immunoglobulin mu binding protein 2	F: GGTGCTTTCACACAAAATGT R: CATTTCCGCATCCATGAATT	117
ORF03066	General transcription factor IIH subunit 3-like	F: GGTTGCGACATAACAGGAGGT R: CTCAAAGGTGGCTCAGGTAG	98
ORF03683	Pangolin	F: GATGGCAATCAGTCAGAGGA R: AGATGTGAGCGAGGAAAGAC	147
ORF05393	Armadillo repeat-containing protein 2	F: TATACTCCACCGCCAAGACT R: TGGTGTAAAAGGTCTGCTGG	126
ORF05808	Homeobox protein abdominal-A	F: CTGAGCCCGAATTCGAACAA R: CTGATGATGATGATGAGCCG	120
ORF06252	Av71 muscle cell intermediate filament	F: ATGTGGGCGGGACTTAAGAA R: CTCGCCCACTTTTTCCTAAG	200
ORF01553	Pro-corazonin preproprotein	F: GATTTCAACTGGTTTGTGCG R: GTCTGCTATCCCACATTTCT	179
ORF02528	Similar to visgun CG16707-PC	F: CCACTCCTGTTACTCCTCCT R: TGGAGTTGTAGGTGGTGGTG	101
ORF02941	RNA binding protein Vera (Insulin-like growth factor 2)	F: TCGGTAGAGACTACGGAACT R: CGAAGATTGGTCTGGGCAAT	171
ORF03027	RNA-binding protein 38-like isoform 1	F: ACCAGGAGGTATCGTACCTT R: CTGCTGCACTAGTGTATGGA	173
ORF03445	Leucine-rich repeat and WD repeat-containing protein 1	F: TGGTGGATCGGTGCAAATGA R: GTGATTCCGACTTTGACTCC	105
ORF04009	UPF0459 protein CG10681-like	F: AAAATACCCCAGAAAGTGAAAGTG R: GCTCTTATCATTGATTCTACATCT	126
ORF04092	Hypoxia-inducible factor 1 alpha	F: CGCCGTCACGTCGAAGATAT R: GAATTCGTCGGAGTCACGGT	120
ORF04093	Hypoxia-inducible factor 1 alpha	F: CCAGGCATGCAGATGAGGAA R: ATGGCCGTACCTTCGGGATT	151

[a] Amp size: Amplified PCR product size in bp.

Supplementary Table 2

List of immune-related genes annotated from the body and head louse genome

Gene Name	Number	Body louse ID	Head louse ID	I
Recognition	**38**			
PGRP	1	PHUM581030	HLORF11059	
GNBP (BGBP)	x[a]			
Fibrinogen-related protein	2	PHUM562660	HLORF10623	
		PHUM500950	HLORF09263	
C-type lectin	9	PHUM467750	HLORF08594	
		PHUM248020	HLORF04183	
		PHUM458550	HLORF08320	
		PHUM390090	HLORF06700	
		PHUM509080	HLORF09507	
		PHUM150830	HLORF02473	
		PHUM151070	HLORF02487	
		PHUM280850	HLORF04774	
		PHUM489310	HLORF08916	
Hemocytin	1	PHUM474690	HLORF08712	
Galectin	3	PHUM402330	HLORF07000	
		PHUM275780	HLORF04708	
		PHUM051550	HLORF00913	
TEP	3	PHUM375050	HLORF06425	
		PHUM289860	HLORF04974	
		PHUM289710	HLORF04971	
Nimrod A	1	PHUM522270	HLORF09845	
Nimrod B	x			
Nimrod C	x			
Draper	1	PHUM049590	HLORF00866	
Dscam	1	PHUM602300	HLORF11620	
Duox	1	PHUM454140	HLORF08183	
Scavenger receptor A	4	PHUM454890	HLORF08199	
		PHUM602700	HLORF11644	
		PHUM066640	HLORF01137	
		PHUM534870	HLORF10080	
Scavenger receptor B	10	PHUM603690	HLORF11674	
		PHUM424210	HLORF07465	
		PHUM569600	HLORF10779	
		PHUM569610	HLORF10780	
		PHUM365540	HLORF06255	
		PHUM569120	HLORF10757	
		PHUM351630	HLORF05958	
		PHUM351640	HLORF05959	

		PHUM365690	HLORF06260
		PHUM563930	HLORF10646
Scavenger receptor C	1	PHUM356530	HLORF06042
Modulator	**22**		
CLIP serine protease	6	PHUM501910	HLORF09292
		PHUM360690	HLORF06138
		PHUM451100	HLORF08107
		PHUM192460	HLORF03113
		PHUM571420	HLORF10835
		PHUM027570	HLORF00524
Serpin	16	PHUM220550	HLORF03612
		PHUM432060	HLORF07679
		PHUM291200	HLORF04991
		PHUM291170	HLORF04988
		PHUM492620	HLORF09064
		PHUM108970	HLORF01738
		PHUM108960	HLORF01738
		PHUM106690	HLORF01704
		PHUM106570	HLORF01703
		PHUM106460	HLORF01702
		PHUM075870	HLORF01225
		PHUM221060	HLORF03625
		PHUM600840	HLORF11548
		PHUM291190	HLORF04990
		PHUM291180	HLORF04989
		PHUM311330	HLORF05347
Toll pathway	**16**		
Spätzle	3	PHUM596260	HLORF11408
		PHUM332090	HLORF05632
		PHUM057390	HLORF00994
Toll	6	PHUM529420	HLORF09977
		PHUM081740	HLORF01330
		PHUM480550	HLORF08823
		PHUM108410	HLORF01728
		PHUM107160	HLORF01719
		PHUM006690	HLORF00161
MyD88	1	PHUM536290	HLORF10117
Tube	1	PHUM194370	HLORF03155
Pelle	1	PHUM518290	HLORF09774
TRAF2	1	PHUM129280	HLORF02091
ECSIT	1	PHUM075600	HLORF01219
Cactus	1	PHUM345810	HLORF05855
Dorsal	1	PHUM534140	HLORF10065

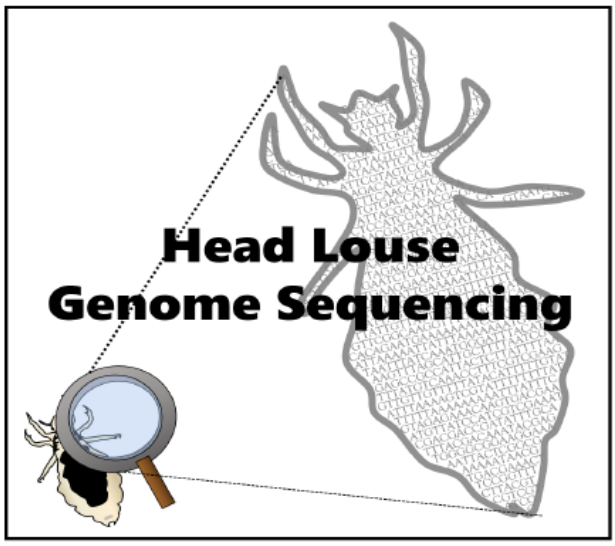
Imd pathway	**6**		
IMD	x		
Dredd	1	PHUM574530	HLORF10867
TAK1	1	PHUM125410	HLORF01932
FADD	x		
Tab2	1	PHUM433350	HLORF04584
IAP2	1	PHUM080100	HLORF01300
IKK beta(IRD5)	1	PHUM605130	HLORF11731
Relish	1	PHUM424590	HLORF07494
JNK pathway	**4**		
Hem	1	PHUM588610	HLORF11246
JNK(Basket)	1	PHUM128040	HLORF02026
Kay	1	PHUM237480	HLORF03977
Jun(jra)	1	PHUM379500	HLORF06548
JAK/STAT pathway	**3**		
Domeless	1	PHUM374950	HLORF06422
JAK	1	PHUM202560	HLORF03268
STAT	1	PHUM335200	HLORF05714
Effector	**4**		
PPO	1	PHUM448900	HLORF08027
Noduler	1	PHUM249370	HLORF04228
Defensin	2	PHUM365700	HLORF12652
		PHUM595870	HLORF11388
Other AMPs	x		
Total	**93**		

[a] not found

Research highlights

- ✓ Head louse genome was analyzed and compared with the reference body louse genome.
- ✓ Genomes of head and body lice possess > 10,790 genes.
- ✓ Gene composition of head and body lice appears to be identical.
- ✓ Nucleotide diversity of genome between head and body lice was 2.2%.
- ✓ *Candidatus* Riesia pediculicola was the primary endosymbiont in both head and body lice.

Graphical Abstract



Commons with body louse genome

- ~110 Mb
- >10,790 genes
- Identical gene composition
- Primary endosymbiont: *Ca.* Riesia pediculicola

Differences from body louse genome

- ave. 2.2% nt sequence divergence
- 5 genes with high *Ka/Ks* ratio

Figure 1