Primates, Lice and Bacteria: Speciation and Genome Evolution in the Symbionts of Hominid Lice

Bret M. Boyd,*1,2 Julie M. Allen,2,3 Nam-Phuong Nguyen,4 Pranjal Vachaspati,5 Zachary S. Quicksall,3 Tandy Warnow,5 Lawrence Mugisha,6,7 Kevin P. Johnson,2 and David L. Reed3

1Department of Entomology, University of Georgia Athens, Athens, GA
2Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, Champaign, IL
3Florida Museum of Natural History, University of Florida, Gainesville, FL
4Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA
5Department of Computer Science and Department of Bioengineering, University of Illinois Urbana-Champaign, Champaign, IL
6Conservation & Ecosystem Health Alliance (CEHA), Kampala, Uganda
7College of Veterinary Medicine, Animal Resources & Biosecurity (COVAB), Makerere University, Kampala, Uganda

*Corresponding author: E-mail: bboyd@uga.edu.

Associate editor: James McInerney
Phylogenetic data have been deposited in TreeBase.org, http://purl.org/phylo/treebase/phylows/study/TB2:S19788 (NCBI Complete Genome accessions CP012838–CP012849 and WGS accession LKAS00000000).

Abstract

Insects with restricted diets rely on symbiotic bacteria to provide essential metabolites missing in their diet. The blood- sucking lice are obligate, host-specific parasites of mammals and are themselves host to symbiotic bacteria. In human lice, these bacterial symbionts supply the lice with B-vitamins. Here, we sequenced the genomes of symbiotic and heritable bacterial of human, chimpanzee, gorilla, and monkey lice and used phylogenomics to investigate their evolutionary relationships. We find that these symbionts have a phylogenetic history reflecting the louse phylogeny, a finding contrary to previous reports of symbiont replacement. Examination of the highly reduced symbiont genomes (0.53–0.57 Mb) reveals much of the genomes are dedicated to vitamin synthesis. This is unchanged in the smallest symbiont genome and one that appears to have been reorganized. Specifically, symbionts from human lice, chimpanzee lice, and gorilla lice carry a small plasmid that encodes synthesis of vitamin B5, a vitamin critical to the bacteria-louse symbiosis. This plasmid is absent in an old world monkey louse symbiont, where this pathway is on its primary chromosome. This suggests the unique genomic configuration brought about by the plasmid is not essential for symbiosis, but once obtained, it has persisted for up to 25 My. We also find evidence that human, chimpanzee, and gorilla louse endosymbionts have lost a pathway for synthesis of vitamin B1, whereas the monkey louse symbiont has retained this pathway. It is unclear whether these changes are adaptive, but they may point to evolutionary responses of louse symbionts to shifts in primate biology.

Key words: Ca. Riesia, endosymbiont replacement, Anoplura, pantothenate, thiamin, plasmid.

Introduction

Insect species with nutritionally incomplete diets (e.g., phloem or blood) often harbor mutualistic bacteria that synthesize missing nutrients (Baumann 2005; Lopez-Sanchez et al. 2009; Manzano-Marin et al. 2015). Many of these symbiotic bacteria (herein endosymbionts) are intracellular, occupy specialized host cells (bacteriocytes; reviewed by Buchner 1965), and are transmitted vertically from mother to offspring (reviewed by Bright and Bulgheresi 2010). Generally, endosymbionts of insects are maintained over evolutionary time-scales and many coevolve with their insect hosts (Clark et al. 2000; Sauer et al. 2000; Thao et al. 2000).

Symbiosis between blood-feeding lice and heritable endosymbionts has been well-studied (Ries 1931, 1932, 1933, 1935; Ascher and Ries 1933; Ries and van Weel 1934; Buchner 1965; Eberle and Mclean 1982, 1983; Sasaki-Fukatsu et al. 2006; Allen et al. 2007, 2009, 2016; Perotti et al. 2007; Boyd and Reed 2012; Boyd et al. 2014, 2016). The most familiar is the human head louse (Pediculus humanus), which has a large visible mass of endosymbiont cells in the abdomen (Hooke 1665; Buchner 1965; Perotti et al. 2007). The role of human louse endosymbionts has been studied experimentally (Puchta 1955) and these endosymbionts supply the lice with essential B-vitamins (Puchta 1955; Perotti et al. 2009). Central to the symbiosis is the synthesis of vitamin B5 (pantothenate; Puchta 1955; Perotti et al. 2009), which is a precursor of coenzyme A and can be transported between the endosymbiont and host cells (Vallari and Rock 1985). Synthesis of vitamin B5 requires both the louse and the endosymbiont, a process known as metabolic complementation (Wilson and Duncan 2015). Here, the louse synthesizes precursors needed for the synthesis of vitamin B5 and the pathway is completed.

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
by the endosymbiont. Vitamin B5 is then available for both the endosymbiont and host.

Human lice belong to the insect suborder Anoplura, more commonly known as the sucking lice (Durden and Musser 1994). This suborder consists of 532 species of lice and each species parasitizes one or a few closely related species of mammals (Durden and Musser 1994). Many of these louse species have been shown to harbor endosymbionts (Ries 1931, 1932, 1933, 1935; Aschner and Ries 1933; Ries and van Weel 1934; Buchner 1965). Because all sucking lice feed on mammal blood, it seems likely that other louse endosymbionts have similar roles of providing B-vitamins. However, it appears that symbioses between lice and bacteria have arisen multiple times in different louse species (Buchner 1965; Hypsa and Krizek 2007; Allen et al. 2009, 2016; Boyd and Reed 2012; Boyd et al. 2016), rather than being derived from a common ancestor.

Focusing on human and primate louse endosymbionts, the classification of these endosymbionts reflects that they may have independent origins. The endosymbionts of lice that parasitize hominids (humans, chimpanzees, and gorillas) are classified into the genus Candidatus Riesia (hereafter Riesia; Sasaki-Fukatsu et al. 2006; Allen et al. 2007, 2009). Riesia appears to have been derived from an Arsenophonus-like bacteria (Allen et al. 2007, 2016; Novakova et al. 2009). 12.95–25 Ma (Allen et al. 2009). This is relatively young compared with similar endosymbionts in different insects that have persisted from 25 to 270 My (reviewed by Gosalbes et al. 2010). However, once established, Riesia appears to have coexisted with their louse hosts (Allen et al. 2007, 2009) and louse speciation was largely driven by coevolution with hominids (Reed et al. 2007; Light and Reed 2009). The lice that parasitize old world monkeys, the sister group of hominid lice (Light and Reed 2009; Light et al. 2010), appear to have an unrelated endosymbiont, Candidatus Puchtella (hereafter Puchtella; studied in red colobus monkeys and macaques thus far and named in macaques; Allen et al. 2009, 2016; Fukatsu et al. 2009). This suggests that there was a replacement of an endosymbiont in the common ancestor of hominid lice.

Riesia species have small genomes (Kirkness et al. 2010; Boyd et al. 2014). The two genomes sequenced thus far from endosymbionts of human and chimpanzee lice consist of a linear chromosome of ~0.57 Mb and a circular plasmid ~8 kb (Kirkness et al. 2010; Boyd et al. 2014). This small genome is typical of many endosymbionts (reviewed by McCutcheon and Moran 2012; Moran and Bennett 2014). However, a unique feature of the Riesia genome is that the genes underlying vitamin B5 synthesis, a pathway key to the symbiosis, are encoded on the plasmid (Kirkness et al. 2010; Boyd et al. 2014). In closely related bacteria, these same genes are encoded the primary chromosome, with genes in two different regions of the chromosome (e.g., Escherichia coli). Encoding essential metabolic functions on a plasmid is unusual (Villasenor et al. 2011), however a similar occurrence has been observed in aphids and their endosymbionts (Buchnera; Baumann et al. 1999; Wernegreen and Moran 2001). The movement of essential functions on to a plasmid may have evolutionary consequences for the endosymbiont. For example, extra steps during cell division may be required to ensure all daughter cells receive the plasmid (e.g., see Summer 1996 for review of segregation instability). However, there may be benefits of this genome organization as well. For example, unifying these genes on a small plasmid brings them into close physical proximity in the genome. This configuration could be beneficial by increasing gene copy number, facilitating increased pantothenate synthesis.

Because Riesia has speciated with hominin primates, this group presents a unique opportunity study the timing and persistence of genomic changes. In this study we sequenced and assembled genomes of endosymbionts from human, chimpanzee, gorilla, and red colobus monkey lice. This includes endosymbionts of three louse ecotypes on humans that have different feeding strategies and ability to transmit disease (Raoult and Roux 1999; Anderson and Anderson 2000; Raout et al. 2006). We examine the characteristics of each endosymbiont genome and focus on their capacity for B-vitamin synthesis; including in which species pantothenate synthesis is located on a plasmid. We then re-evaluate the phylogenetic placement of louse endosymbionts, Riesia and Puchtella using genomic data and compare with past phylogenetic studies of these endosymbionts that were based on only one or two genes (Sasaki-Fukatsu et al. 2006; Allen et al. 2007, 2016; Fukatsu et al. 2009). Finally, we evaluate potential symbions replacement.

Results

Species Relationships

We investigated the species relationships of primate and human louse endosymbionts using phylogenomic methods. First, we identified 177 single-copy orthologous protein-coding genes across 38 eubacterial taxa (37 Enterobacteriaceae and one outgroup taxon), including 8 louse endosymbionts. Third codon positions, which had higher frequency of AT bases in endosymbionts, were excluded from the analysis. Using two different software packages, two separate species trees were then estimated using coalescent-gene tree summary methods. In both cases, all of the louse endosymbionts included in this study were found to be monophyletic (fig. 1; supplementary figs. S2 and S3, Supplementary Material online). Endosymbionts of tsetse flies (Wigglesworthia) were identified as the closest relative of louse endosymbionts in our data set. Support for the monophyly of louse endosymbionts and bipartitions within the louse endosymbiont clade were high (0.75–1.0). Next, we again estimated the species relationships based on simultaneous analysis of all genes as a single data matrix. The concatenation of all gene alignments provided 128,006 sites for this analysis, and a Maximum Likelihood (ML) tree was estimated from the concatenated alignment. The resulting ML species tree agreed with coalescent analysis finding that louse endosymbionts were monophyletic and sister to Wigglesworthia (fig. 1). Support for this relationship was again high, 94% bootstrap for monophyly of louse endosymbionts, and 100% bootstrap supporting bipartitions within the louse endosymbiont clade. While it was previously suggested that
**Fig. 1.** Comparison of phylogenetic trees of gamma-proteobacteria including louse endosymbionts obtained from different phylogenomic methods. Left is ML tree based on concatenation of 177 orthologous protein-coding genes. Support is presented as percent of times each bipartition was recovered from 100 bootstrap replicates. Middle and right are coalescent trees (ASTRAL and ASTRID) based on analyses of 177 individual ML gene trees. Red = Riesia; light blue = Puchtella, and purple = Arsenophonus.
Riesia was closely related to Arsenophonus, we failed to recover that relationship in any phylogenetic analysis.

Within the louse endosymbiont tree, we find phylogenetic patterns of the endosymbionts are similar to the lice they inhabit. The endosymbiont Puchtella, from the lice of red colobus monkeys, is sister to the hominid louse endosymbionts, which is consistent with the louse phylogeny described by Light and Reed (2009) and Allen et al. (2017). Also, inclusion of the gorilla louse endosymbiont shows that it is closely related to the human pubic louse endosymbionts. This was expected as gorilla lice and human pubic lice are closely related (a pattern resulting from lice switching from gorillas to humans; Reed et al. 2007; Light and Reed 2009). Also as expected, human and chimpanzee louse endosymbionts are sister to each other. Within human louse endosymbionts, again the tree reflects phylogenetic patterns seen in the lice (Reed et al. 2004). Within human head and body lice there are three distinct clades of lice identified by their mitochondrial haplotypes, potentially resulting from exchange of lice between modern and archaic humans (Reed et al. 2004). We examined endosymbionts from human lice belonging to two of these clades; “Clade A” which is found worldwide and includes both human head and body lice and “Clade B” which is found primarily in western Europe and the New World and only includes head lice (Reed et al. 2004). Again the phylogeny of the endosymbionts reflect their host’s mitochondrial history, with Clade A louse endosymbionts being most closely related to each other and a Clade B louse endosymbiont sister to Clade A strains. Unlike human louse endosymbionts, we see no divergence in chimpanzee louse endosymbionts. This is despite known population structure and limited gene flow in chimpanzees (Becquet et al. 2007). Perhaps additional sampling will reveal structure within this species.

Endosymbiont Genomes

We completely or partially sequenced the genomes of six new strains of louse endosymbionts (Puchtella and Riesia species; table 1). Similar to the two published Riesia genomes from human body and chimpanzee lice, the newly sequenced Riesia genomes (from human head lice, human pubic lice, chimpanzee lice, and gorilla lice) consist of a primary chromosome ranging from 0.529–0.574 Mb in length and a small circular plasmid 5.2–7.8 kb (table 2 and fig. 2). The previously sequenced endosymbiont genome from a human louse was the USDA strain of Riesia pediculicola, which is derived from an inbred strain of lice adapted to feed on rabbits in the laboratory (Kirkness et al. 2010). Here, we find the genome of the endosymbiont derived from these rabbit-adapted lice is similar to the genomes of endosymbiont in wild populations of human head lice (fig. 2). The genomes of A clade human head lice and body lice were nearly identical (99.9% of bases were the same in both genomes) with no large insertions or deletions (indels). There was slightly more variation between endosymbiont genomes of head lice, than between the endosymbionts genomes of the body louse and the most closely related head louse endosymbiont. Compared with the body louse endosymbiont, the B clade endosymbiont genome is more divergent (97.8% of bases were identical) and has three
biont genomes from all lice were similar. Much of the coding prediction in each endosymbiont. Functionally, the endosymbiont of the gorilla louse was slightly smaller than the human body louse genome, however we identified a similar genome (from the endosymbiont of the gorilla louse) and in Puchtella, which has seen extensive reorganization of its genome when compared with Riesia species.

Synthesis of B-vitamins is likely the primary role of endosymbionts in blood sucking lice. We compared B-vitamin biosynthesis in Puchtella with Riesia (focusing the human body louse endosymbiont as gene content was similar across strains; fig. 4). Both species possess the de novo pathway for synthesis of pantothenate from exogenous substrates. They can also synthesize Coenzyme A (CoA) from pantothenate. Synthesis of folic acid, riboflavin, pyrodoxine phosphate, and biotin also appear intact. The only difference between the two genera is that Riesia appears to be missing the gene encoding 4-erythrose 4-phosphate dehydratase (epd) required for pyridoxine biosynthesis, while Puchtella retains this function. However, because the remainder of the pyridoxine synthesis pathway is intact in Riesia, the missing epd encoding gene could be due to an assembly or annotation error. Metabolism of nicotinamide was different in Riesia and Puchtella with each using different starting compound (fig. 4).

Finally, we found that Puchtella with each using different starting compound (fig. 4).

In Riesia, proteins involved in pantothenate synthesis (a pathway critical to symbiosis with lice) were encoded on the plasmid (from the endosymbiont of the gorilla louse) and in Puchtella, which has seen extensive reorganization of its genome when compared with Riesia species.

Table 2. Genome Assembly Summary of Louse Endosymbionts.

<table>
<thead>
<tr>
<th>Taxonomy</th>
<th>Seq. Technology</th>
<th>Seq. Depth</th>
<th>SD Depth</th>
<th>Genome Length</th>
<th>Genome %GC</th>
<th>Plasmid Length</th>
<th>Plasmid %GC</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riesia USDA</td>
<td>Sanger</td>
<td>NA</td>
<td>NA</td>
<td>574,390</td>
<td>28.5</td>
<td>7737</td>
<td>35.2</td>
<td>557</td>
</tr>
<tr>
<td>Riesia HHAC</td>
<td>Illumina HiSeq 2000</td>
<td>22.9.2</td>
<td>51.5</td>
<td>574,389</td>
<td>28.5</td>
<td>7737</td>
<td>35.3</td>
<td>566</td>
</tr>
<tr>
<td>Riesia HHAN</td>
<td>Illumina HiSeq 2500</td>
<td>2.3</td>
<td>1.5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Riesia HHHB</td>
<td>Illumina HiSeq 2000</td>
<td>424.6</td>
<td>107</td>
<td>574,386</td>
<td>28.5</td>
<td>7737</td>
<td>35.9</td>
<td>573</td>
</tr>
<tr>
<td>Riesia PTSU</td>
<td>Illumina HiSeq 2000</td>
<td>147.2</td>
<td>36.4</td>
<td>566,667</td>
<td>31.6</td>
<td>5197</td>
<td>37</td>
<td>NA</td>
</tr>
<tr>
<td>Riesia PTSK</td>
<td>Illumina HiSeq 2500</td>
<td>164.1</td>
<td>64.4</td>
<td>566,667</td>
<td>31.6</td>
<td>5197</td>
<td>37</td>
<td>594</td>
</tr>
<tr>
<td>Riesia HPNS</td>
<td>Illumina HiSeq 2500</td>
<td>9</td>
<td>99.7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Riesia GBBU</td>
<td>Illumina HiSeq 2500</td>
<td>83</td>
<td>39.1</td>
<td>528,693</td>
<td>25</td>
<td>5651</td>
<td>29</td>
<td>476</td>
</tr>
<tr>
<td>Puchtella</td>
<td>Illumina HiSeq 2000</td>
<td>396.4</td>
<td>71.3</td>
<td>558,106</td>
<td>24.2</td>
<td>NA</td>
<td>564</td>
<td>NA</td>
</tr>
</tbody>
</table>

Note.—SD, standard deviation; Seq. sequencing.
may have some important, but as-yet undetermined role. There is a small repeated region between the chaperonin protein and the \textit{panE} gene in plasmids recovered from human louse endosymbionts. This repeat region is missing in other Riesia genomes.

**Gene Tree–Species Tree Conflict**

A phylogenetic analysis of the genes on the plasmid and their orthologs in other taxa was conducted. An ML tree based on concatenation of plasmid-based genes produced a topology different from that of the species trees described above (genes in the Riesia plasmid appear mostly closely related to orthologs in \textit{Arsenophonus}; supplementary fig. S4, Supplementary Material online). However, bootstrap support for this plasmid tree was low and therefore difficult to interpret. The plasmid tree was based on only four genes and it is possible that gene trees/species tree conflict or a lack of informative bases within each gene is driving this alternate typology. To better understand how gene trees contribute to phylogenetic signal in isolation, we compared individual gene trees with the ML species tree that was based on 177 genes. When gene trees and the species tree were collapsed leaving only bipartitions with greater than 50\% bootstrap support, a mean of 20\% of bipartitions in the gene trees conflicted with the species tree (mean = 0.20, Q1 = 0.13, Q3 = 0.27; range = 0.0–0.71; supplementary fig. S5, Supplementary Material online). However, when collapsing bipartitions at higher values, the conflict decreased considerably (collapsing at 75\% bootstrap
mean = 0.06, Q1 = 0.0, Q3 = 0.14; range = 0.0–0.5; and at 95% bootstrap mean = 0.03, Q1 = 0.0, Q3 = 0.0; range = 0.0–0.5; supplementary fig. S5, Supplementary Material online). These results show that conflict between individual gene trees and the species tree is pervasive, but these deviations are not well supported. Those genes on the plasmid showed a similar pattern of conflict that is not supported by bootstrap replicates (50%, mean = 0.31, Q1 = 0.26, Q3 = 0.39; range = 0.16–0.4; 75%, mean = 0.12, Q3 = 0.18; range = 0.07–0.18; 95%, mean = 0.02, Q1 = 0.0, Q3 = 0.02; range = 0.0–0.08; supplementary fig. S5, Supplementary Material online). These comparisons suggest that lack of support limits phylogenetic tree estimation and that conflict between the species tree and individual gene trees is not unique to the plasmid genes.

Previous phylogenetic studies examining primate louse endosymbionts have provided evidence for frequent endosymbiont replacement (Sasaki-Fukatsu et al. 2006; Allen et al. 2007, 2016; Hypsa and Krizek 2007; Fukatsu et al. 2009). It had been suggested that the plasmid could move during such replacements (Kirkness et al. 2010) through horizontal exchange (HGT). To test for evidence of HGT, we determined if plasmid-based genes had a phylogenetic topology that significantly differed from the presumed species tree when compared with genes found on the primary chromosome (where HGT is not expected). A Welch two-sampled t-test failed to reject the null hypothesis that the plasmid genes have a conflict that is significantly higher than genes on the primary chromosome as would be expected under HGT. Results of t-test are as follows when collapsing tree bipartitions at 50% bootstrap t(3) = −1.9095, P > 0.05; collapsing at 75% bootstrap t(3) = −1.8504, P > 0.05; and collapsing at 90% bootstrap t(3) = −0.3851, P > 0.05. These results indicate that phylogenetic trees estimated from plasmid-based genes are statistically compatible to phylogenetic trees estimated from other genes. Therefore we find no evidence of HGT.

**Discussion**

Here we examine the phylogenetic relationships and genome structure of diverse primate louse endosymbionts. Our results agree with previous phylogenetic studies that suggested endosymbionts have cospeciated with human and chimpanzee which diverged 5.4 Ma (Allen et al. 2009) and gorilla lice which diverged 7.4–17.4 Ma (Allen et al. 2007, 2009). We also describe new evidence that endosymbionts co-diverged along with their gorilla and human pubic louse hosts, which diverged following a host switch ~3–4 Ma (Reed et al. 2007). We also find that endosymbionts in human head and body lice have similar phylogenetic patterns to human louse mitochondria. The endosymbionts of human head and body lice, both belonging to the louse mitochondrial clade A have nearly identical genomes. This is not surprising as human body lice diverged from head lice ~80–100 ka, after humans began wearing clothes (Toups et al. 2011). Therefore, the lab-strain body lice have endosymbionts that are representative of endosymbionts from A clade head lice, a common strain of lice found world-wide (Reed et al. 2004). However, we did find there was more variation between human head louse endosymbiont genomes that corresponded to different louse mitochondrial clades (clades A and B). Reed et al. (2004) found evidence that these louse strains diverged up to 1.18 Ma. They proposed that these lice diverged on extinct humans and were acquired by modern humans via a host switch in Europe. Our finding that the genomes of endosymbionts of head lice are more divergent than A clade head and body lice fits with what is known about louse evolution.
However, our findings diverge from previous studies in two ways. First, we find no evidence that the human and other hominid louse endosymbionts, Riesia, are closely related to the genus Arsenophonus. Instead, we find they are closely related to the endosymbionts of a primate louse, Puchtella, which is found in the lice of red colobus monkeys (fig. 1; Boyd et al. doi:10.1093/molbev/msx117 MBE 1750).
supplementary figs. S2 and S3, Supplementary Material online). Second, the close relationships of Riesia and Puchtella suggests there was no endosymbiont replacement in the common ancestor of hominid lice; instead the phylogenetic pattern of the louse endosymbionts reflects the species history of the lice themselves (fig. 6; see Light and Reed 2009; Light et al. 2010; and Allen et al. 2017 for the louse phylogeny). Our results support a scenario of repeated cospeciation of endosymbionts and lice. These findings also push back the date of the association of these endosymbionts with lice. Based on host divergence information, it appears hominid lice and old world monkey lice diverged between 20 and 25 Ma (Light and Reed 2009; Light et al. 2010), suggesting the louse-endosymbiont association is at least that old. Essentially, this constrains the association between lice and this clade of endosymbionts to the upper end of a previously reported data range of 12.9–25 Ma (Allen et al. 2009). Future work should critically evaluate the relationship of the louse endosymbionts studied here and endosymbionts from other old world monkey lice not included in this study. Possibly the association between Riesia species and lice has persisted for more than 25 My.

Correctly placing insect-endosymbionts, including endosymbionts of lice, into the phylogenetic tree of Enterobacteriaceae has been the focus of many studies (e.g., Herbeck et al. 2005; Williams et al. 2010; Sach et al., 2011, 2014; Allen et al. 2016). The reason that previous phylogenetic studies based on only one or two genes arrived at a different result for the placement of Riesia and Puchtella (e.g., Hypsa and Krizek 2007; Allen et al. 2007, 2009, 2016; Fukatsu et al. 2009; Novakova et al. 2009 using 16S rRNA or groEL), may be due to the use of single gene trees that show a history incongruent with the species history, or that provide an insufficient amount of phylogenetically informative data. Our results show there is often conflicting signals between gene trees, but support for this conflict is low. Therefore, phylogenies built on one or two genes may not be sufficient to resolve the louse endosymbiont evolutionary tree. Phylogenetic studies seeking to more broadly evaluate polyphyletic endosymbionts in sucking lice should be based on many genes.

Additional issues arise beyond gene tree–species tree conflict. The endosymbiont of the human louse, Riesia, has a DNA substitution rate higher than most other endosymbionts (Allen et al. 2009). Like in other endosymbionts, mutation favors AT bases (Moran 2002; Van Leuven and McCutcheon 2011). This high rate of substitution and AT bias may cause endosymbionts to artificially cluster together in a phylogenetic tree, despite them being derived from independent lineages (long-branch attraction; Felsenstein 1978). One approach that provides opportunity to overcome some phylogenetic issues is implementing a nonhomogeneous model of sequence evolution (Herbeck et al. 2005). Unfortunately, this approach cannot yet be implemented with large phylogenomic data sets and relies on limited methods for assessing alternative trees (nearest neighbor interchange). To attempt to limit the effect of AT base bias, we eliminated third codon positions. This position showed the greatest bias in endosymbionts compared with other bacteria. In an attempt to detect methodological shortcomings of tree building with Enterobacteriaceae, we used two different methods of summarizing the species tree. First, using the concatenation method, all the sequence data were analyzed as a single matrix to estimate a species tree. This approach was appealing as bacterial genes are shorter than eukaryotic genes and individual genes may provide limited phylogenetically informative variants. The second approach summarized species trees from gene trees. This approach is appealing as different classes of genes might experience different rates of substitution or mutation biases. Therefore, each gene tree can contribute to the resulting species tree equally. These methods both agree that Wigglesworthia and louse symbionts are the closest relatives in our analysis. It seems unlikely that Wigglesworthia was directly exchanged between insects. One possibility is that Wigglesworthia and louse symbionts were both derived from a closely related “progenitor strain” of bacteria. This scenario would be like Sodalis species, which have been detected in distantly related insects, including lice (Fukatsu et al. 2007; Smith et al. 2013; Boyd et al. 2016).

The genomes of all louse symbionts sampled are small and much of the genome is devoted to vitamin synthesis. In two instances, we noted the loss of metabolic function potentially facilitated by genome reduction and reorganization. These losses largely impacted fatty acid and amino acid metabolism, while leaving vitamin metabolism intact. The synthesis of B-vitamins is important to the symbiosis between endosymbionts and lice. A comparison of Riesia and Puchtella found that both have similar B-vitamin synthesis capabilities. One major difference is that Puchtella retains the pathway for synthesis of thiamin while Riesia does not. The loss of thiamin biosynthesis was also noted in the Sodalis-like endosymbiont of seal lice (Boyd et al. 2016). In this Sodalis-like endosymbiont, pseudogenes indicated the recent loss of this pathway. It is unclear if the retention of thiamin synthesis in Puchtella has any biological relevance. Red colobus monkeys feed on a variety of resources, including plants (leaves and flowers), li-chens, invertebrates, and soil (Struhsaker 2010). However, young leaves constituted the majority of their diet (Struhsaker 2010). It is possible that this leaf heavy diet is low in thiamin and the lice are also deprived of this essential vitamin. Therefore, their parasitic louse may require thiamin from its endosymbionts, while human and chimpanzee lice parasitize hosts with more complex diets that contain thia-min. Hence their endosymbionts of human lice scavenge thiamin rather than synthesizing thiamin (Boyd et al. 2014).

We find that human, chimpanzee, and gorilla louse endosymbionts (Riesia species) carry the small plasmid needed for vitamin B5 synthesis as part of their genomes (table 2 and fig. 5). However, the sister taxon, Puchtella, endosymbiont from the red colobus monkey, does not. Instead, the genes for vitamin B5 synthesis are located on its primary chromosome, a configuration more typical of Eubacteria. It seems likely that the plasmid arose through genome re-organization in the common ancestor of hominid louse endosymbionts. When comparing the genomes of Riesia and Puchtella, there appears to have been movement of genomic regions (fig. 3). This
reorganization may have facilitated the origins of the plasmid. The lack of the plasmid in the red colobus monkey endosymbionts, Puchetella, also suggests that this chromosomal structure is not essential for louse endosymbionts. A recent study of an endosymbiont of another mammalian sucking louse (Boyd et al. 2016) also found this pathway was not plasmid based, but has a similar arrangement to Puchetella. Therefore, movement of these genes into close physical proximity on the genome is not essential; however, it may be advantageous, because it has been conserved through multiple rounds of speciation events.

In many ways, this system appears analogous to the movement of amino acid synthesis genes onto plasmids in the endosymbiont of aphids (Buchnera; Baumann et al. 1999; Wernegreen and Moran 2001). Here the rate-limiting steps in tryptophan and leucine synthesis are encoded on plasmids and these plasmids have been maintained over evolutionary time scales (Rouhbakhsh et al. 1996; Wernegreen and Moran 2001). Wernegreen and Moran (2001) noted there was a slight increase in the substitution rate in genes on plasmids in Buchnera. We compared the pairwise distance between genes on the plasmid and in the linear chromosome of Riesia and only see slight greater pairwise distance in plasmid-based genes (supplementary table S2, Supplementary Material online). It seems likely that the plasmid-based genes in louse endosymbionts are not experiencing a significantly higher substitution rate.

Our results are a departure from previous findings and support monophyly of human and primate louse endosymbionts instead of endosymbiont replacement. Previous studies have found that Riesia species were derived from an Arsenophonus-like bacteria ~ 20–25 Ma (Allen et al. 2009; Novakova et al. 2009). Unlike Riesia, which have genomes ranging from 0.53 to 0.58 Mb in size, Arsenophorus endosymbionts possess larger genomes. Sequenced genomes include the 3.67 Mb genome of A. nasoinae (Wilkes et al. 2010) and 2.95 Mb genome of Arsenophorus sp. (Xue et al. 2014). This would have suggested there was a rapid reduction of the Riesia genome shortly after it colonized lice. Instead endosymbionts of lice might be more ancient and have undergone more long-term genome reductions. This system is biologically interesting, because louse, and hence endosymbiont evolution, are uniquely tied to primate and human evolution. Therefore, shifts in primate biology might impact endosymbiont genome evolution and loss of metabolic functionality (e.g., loss or retention of thiamin synthesis). Phylogenomics and methodological advances will inevitably provide additional insights into this group.

**Materials and Methods**

**Sample Collection**

Endosymbionts are intracellular and obligate associates of insects. To sequence the genomes of different louse endosymbionts, we first collected lice from humans and primates. This included lice from red colobus monkeys, chimpanzees, mountain gorillas, and humans (table 1). Chimpanzees have structured populations with limited gene flow (Becquet et al. 2007). Human head lice have divergent mitochondrial haplogroups (Ashfaq et al. 2015). Therefore, we expected divergent strains of endosymbionts and identified likely candidate lice using the mitochondrial cytochrome oxidase subunit1 (COI; supplementary methods, Supplementary Material online). Because mitochondria and endosymbionts are both maternally inherited, mitochondria were treated as proxies for identifying divergent endosymbiont groups. Each endosymbiont used in sequencing was assigned a strain name based on host association and collection site (supplementary methods, Supplementary Material online).

**DNA Extraction and Sequencing**

Total gDNA was extracted from chimpanzee, gorilla, and human pubic lice using the Zymo Genomic DNA-Tissue MicroPrep kit following the manufacturer’s instructions except that lice were manually macerated prior to incubation with proteinase K. Total gDNA from red colobus monkey lice was extracted using a phenol–chloroform method following Boyd et al. (2014). DNA extracts from the Netherlands human head lice were provided by Ascunce (see Ascunce et al. 2013 for collection methods).

gDNA from each louse was sonicated using the Covaris M220 instrument to an average fragment size of 300–450 bp (actual range was 200–600 bp). The sheared gDNA was prepared for next-generation sequencing using TruSeq DNAseq or Kapa Library preparation kits. The resulting library was sequenced on one-half lane of Illumina HiSeq2000 or 2500 using the TruSeq SBS sequencing kit v.1-2 for 161 cycles. All samples were sequenced paired-end, with 100 or 160 bp reads. Fastq files were produced using Casava v.1.8.2. Sequence data for the Uganda chimpanzee louse were obtained from the Genbank Short Read Archive (accession...
SRX390495; see Boyd et al. 2014 and Johnson et al. 2014 for complete sequencing methods).

**Genome Assembly**

The gDNA libraries contained sequence data from the lice, any nonendosymbiont bacteria, and the endosymbionts. We used a combination of approaches to isolate the endosymbiont data and assemble their genomes. This included reference based and de novo assembly methods. To reconstruct the human louse endosymbiont genomes, we used a reference-guided approach. We first aligned Illumina paired-end reads to the Riesia pediculicola str. USDA genome sequence (gi295698239 and gi292493920) using bowtie2 v.2.1.0 (local option, default sensitivity, results exported as SAM file; Langmead and Salzberg 2012). Resulting SAM files were converted to BAM and sorted using SAMtools v.0.1.19 view and sort functions (Li et al. 2009). We then visualized the sorted BAM files in Geneious v.7.1.7 (www.geneious.com) and searched for evidence of large insertions or deletions (indels) and to check for assembly errors. To make an accurate assembly, it was essential to check for indels as it was suspected that these strains may have diverged more than 1 Ma, ample time for indels to have arisen. No evidence of indels was found in human louse endosymbionts from Cambodia and Netherlands. These endosymbionts came from lice that had the same COI haplotype as the source for the reference genome. However, we did find evidence for indels in the human louse endosymbiont from Honduras when compared with the USDA genome (table 1). This Honduras sample came from a louse with a different COI haplotype. Therefore, to obtain a more accurate assembly of this divergent strain, we used the aTRAM software (Johnson et al. 2013; Allen et al. 2015) to build small-targeted de novo contigs around potential indels. The resulting aTRAM contigs were aligned to the str. USDA genome and the genome sequence was modified to include insertions and remove larger deletions. This modified sequence served as a reference and we aligned reads to it using bowtie2 (end-to-end, default sensitivity, results exported as a SAM file). A consensus genome sequence was then called using vcfutils.pl (script distributed as part of BCF/samtools package) from all final BAM files using this modified reference.

To reconstruct the genomes of chimpanzee louse endosymbionts, we downloaded the Riesia pediculicidaefi str. PTSK genome assembly (gi746672278) and the Illumina read library used in that assembly (SRX390495). The original str. PTSK genome consisted of five genomic scaffolds that made up the primary linear chromosome. To improve this genome build, we used the aTRAM software to build small contigs that overlapped the end existing str. PTSK contigs. We then aligned the new aTRAM contigs to the existing contigs to assemble the primary chromosome into one continuous genome sequence. Next, we aligned the Illumina paired-end reads to the corrected genome sequence using Bowtie2 (end-to-end, default sensitivity, results exported as SAM file). The data were converted to a BAM file and sorted using samtools and the genome consensus sequence called using vcfutils.pl. This corrected genome sequence served as the reference for a reference guided assembly of the Riesia pediculicidaefi str. PTSK genome. Illumina paired-end reads were mapped to the updated PTSK genome sequence using bowtie2 (end-to-end, default sensitivity, results exported as SAM file). The SAM file was converted to a BAM file and sorted using samtools and the genome consensus sequence called using vcfutils.

There was no close reference genome available for the gorilla or the human pubic louse endosymbiont genomes; therefore, we took a combined de novo and reference guided approach. First, we trimmed low quality bases from the Illumina 2500 sequence data using Trimmomatic v.032 (Bolger et al. 2014) by removing bases with a phred score < 30 from the 3’ end of the read for both libraries. After trimming, any reads shorter than 75 bp and their mate pairs were removed. The quality trimmed reads were then de novo assembled into contigs and scaffolds using abyss-pe v.1.5.2 (k = 64). We then used BLASTn v.2.2.28+ to identify assembled contigs belonging to the endosymbiont genome (str. USDA genome served as the target for BLASTn). Next, we used LAST v.531 to find the assembled plasmid by aligning the conserved protein coding sequences to the contigs library (str. USDA pPAN plasmid served as the target sequence). As with the str. PTSK genome, we used aTRAM to build targeted small de novo assemblies to join the endosymbiont contig ends. Contigs were aligned in Geneious and a reference sequence called. We then aligned the Illumina paired-end reads to the draft genome sequence using bowtie2 (end-to-end, default sensitivity). We then examined the resulting read alignments in BAM format to identify assembly errors. The final consensus genome sequence was called using vcfutils.pl. The pubic louse endosymbiont genome was expected to be similar to the gorilla louse endosymbiont. Therefore, we aligned the pubic louse endosymbiont reads to the gorilla louse endosymbiont genome using bowtie2 (local, default sensitivity). The genome sequence was called using vcfutils.pl. Overall, sequencing depth for the pubic louse endosymbiont genome was low and we could not identify indels like we did with human head louse endosymbionts. Therefore we skipped this step for this genome.

In all instances where a plasmid was found in de novo or reference guided assemblies, we used read associations to verify the chromosome was circular. We did this by retrieving BAM files from the assemblies and isolating the plasmid-contig and its reads. The BAM was then visualized in Geneious and we manually verified that read pairs supported circularization of the plasmid by overlapping the end of the contig. Finally, each of the Riesia genomes were annotated using the RAST pipeline (Overbeek et al. 2005; Aziz et al. 2008).

The genome of the red colobus monkey louse endosymbiont was largely unknown. Therefore, we followed the same method described by Boyd et al. (2014) to assemble and annotate the original chimpanzee louse endosymbiont genome (Riesia pediculicidaefi str. PTSK) genome when it was unknown. This method used a de novo assembly of all reads into contigs, BLAST was used to identify contigs belonging to the endosymbiont genome (target consisted of a custom
database of representative bacterial genomes), the genome was re-assembled in isolation with only reads belonging to the genome, and annotation of the genome in the RAST pipeline.

To compare the annotated genes in louse endosymbionts we used reciprocal best tBLASTx 2.5.0+ to identify shared genes (scripts for filtering blast results obtained from http://goolg/csx730). However, this only provides a summary of potential shared genes without regard to functionality of genes. To better understand the metabolic capacity for each genome we used comparison tools in SEED (Overbeek et al. 2005) to compare each endosymbiont genome to that of the genome of the endosymbiont from the human body louse. Total numbers of genes belonging to different functional categories were summarized for each genome and then shared and unique functions were identified using the “compare function” tools within SEED.

Phylogenetic Database Construction

We searched for Enterobacteriaceae with fully sequenced genomes and selected taxa that were representative of the class along with an outgroup taxa from neighboring class, Pasteurellaceae. We then downloaded all protein coding genes as DNA sequences. To find potential orthologs in downloaded gene sequences, we found bidirectional best hits between the genes. First, tBLASTx v.2.2.28+ was used to find hits between genes in all possible pair-wise comparisons of all taxa (default parameters, outformat -6). Second, BLAST hits were filtered to find only the bidirectional best hits between every possible taxa set (scripts for filtering blast results obtained from http://goolg/csx730). Interpreting bidirectional best hits as edges in a network, we grouped genes into separate interconnected networks. Network analysis was done using Cytoscape v.3.2.1 (Shannon et al. 2003), using the “network analysis” and “find interconnected subnetwork” tools. Node lists for all subnetworks that possessed between 45 and 10 nodes were exported from Cytoscape with 195 total subnets recovered. To remove paralogs, we then removed any taxa that were represented by two or more nodes in a network. Any subnets with fewer than ten nodes after paralog removal were then removed leaving 177 subnets that are herein treated as groups of orthologous protein-coding genes. Networks containing genes found on the Riesia plasmid were manually annotated to ensure accuracy. To do this, we aligned based on best hits to the Riesia genes, and alignment of the Riesia plasmid were manually annotated to ensure accuracy.

To compare the annotated genes in louse endosymbionts we used reciprocal best tBLASTx 2.5.0+ to identify shared genes (scripts for filtering blast results obtained from http://goolg/csx730). However, this only provides a summary of potential shared genes without regard to functionality of genes. To better understand the metabolic capacity for each genome we used comparison tools in SEED (Overbeek et al. 2005) to compare each endosymbiont genome to that of the genome of the endosymbiont from the human body louse. Total numbers of genes belonging to different functional categories were summarized for each genome and then shared and unique functions were identified using the “compare function” tools within SEED.

Phylogenetic Database Construction

We searched for Enterobacteriaceae with fully sequenced genomes and selected taxa that were representative of the class along with an outgroup taxa from neighboring class, Pasteurellaceae. We then downloaded all protein coding genes as DNA sequences. To find potential orthologs in downloaded gene sequences, we found bidirectional best hits between the genes. First, tBLASTx v.2.2.28+ was used to find hits between genes in all possible pair-wise comparisons of all taxa (default parameters, outformat -6). Second, BLAST hits were filtered to find only the bidirectional best hits between every possible taxa set (scripts for filtering blast results obtained from http://goolg/csx730). Interpreting bidirectional best hits as edges in a network, we grouped genes into separate interconnected networks. Network analysis was done using Cytoscape v.3.2.1 (Shannon et al. 2003), using the “network analysis” and “find interconnected subnetwork” tools. Node lists for all subnetworks that possessed between 45 and 10 nodes were exported from Cytoscape with 195 total subnets recovered. To remove paralogs, we then removed any taxa that were represented by two or more nodes in a network. Any subnets with fewer than ten nodes after paralog removal were then removed leaving 177 subnets that are herein treated as groups of orthologous protein-coding genes. Networks containing genes found on the Riesia plasmid were manually annotated to ensure accuracy. To do this, we aligned based on best hits to the Riesia genes, and alignment of the Riesia plasmid were manually annotated to ensure accuracy.

Species Tree Estimation

All 177 single copy orthologs identified in the previous section were used in determining a potential species tree for louse endosymbionts and representative Enterobacteriaceae. Sequences for each gene were translated in Geneious using bacterial translation table 11. Amino acid sequences were then aligned using UPP v.2.0 (Nguyen et al. 2015) and the aligned sequences back translated to nucleotide sequences. Base composition in the first, second, and third positions were plotted. The plot revealed that the %GC content of the third codon position was higher in endosymbiont species when compared with first and second codon positions (Supplementary fig. S1, Supplementary Material online). This bias would violate the model assumptions of the GTR + gamma substitution model used in ML tree reconstruction below, therefore the third codon positions were omitted from the gene alignments.

We estimated species trees using 1) two coalescent approaches that estimate species trees from gene trees and 2) a concatenation method where a ML tree is estimated from single data matrix of concatenated sequences. For the coalescent approach we first built individual gene trees for all 177 ortholog sets. Each gene tree was estimated using RaxML (v.8.1.3; Stamatakis 2014) under a GTR + gamma model. The species tree was then estimated from these gene trees using ASTRAL (v.4.9.7; Mirarab et al. 2014a; Mirarab and Warnow 2015) and ASTRID (v.1.4; Vachaspati and Warnow 2015). Support for branches in either tree was computed using local posterior probabilities based on gene tree quartet frequencies (Sayyari and Mirarab 2016).

For the concatenation approach, all aligned orthologs were concatenated into a single matrix. In order to account for substitution rate heterogeneity between genes, we performed a partitioned ML analysis by first estimating the GTR + gamma model parameters for each codon position of each gene. We then performed a principal coordinates analysis of the rate parameters and clustered the codon positions using k-means clustering. We found that seven clusters explained most of the variation in the data. The alignment was partitioned according to the clusters and RaxML was used to find the ML tree using a GTR + gamma model, with an initial starting tree based upon a ML tree estimated using FastTree-2 under GTR + cat (v 2.1.7; Price et al. 2010). Support for branches in the concatenation tree was based on 100 bootstrap replicate trees.
Origins of a Small Plasmid

We further investigate the phylogeny of a small plasmid found in many louse endosymbionts to determine if it was congruent with the species tree. First we built an ML phylogeny of the plasmid based on a concatenation of four genes present on the plasmid and their orthologs in other bacterial species. This tree produced a topology incongruent with any of our estimated species trees with regard to louse endosymbionts, however, this tree was based on only four genes (2,176 sites total). This limited sampling of genes might bias the estimated tree topology.

To explore how decreasing gene sampling may have affected this tree, we compared each gene tree with the presumed species tree (the concatenation tree was selected as the species tree for this comparison because both gene trees and the concatenation tree used bootstraps as a measure of bipartition support). For the gene trees, used the ML gene trees described previously (see section on Coalescent species tree estimation). Each gene tree was examined if to determine if all 38 taxa were present. If one or more taxa were missing in the gene tree, these missing species were trimmed from the species tree for that comparison. By doing this we are comparing two trees with the same tips. Next we collapsed unsupported bipartitions in both the gene and species trees. Bootstrap values were used to identify unsupported bipartitions. Finally, we determined the fraction of supported bipartitions in the gene tree that were in conflict with the species tree using scripts described in Mirarab et al. (2014b). The process of collapsing unsupported bipartitions and comparing the trees was done three times using 50, 75, and 95% bootstraps as cutoffs for unsupported bipartitions.

The resulting gene tree–species tree comparisons yielded 0–1 score of similarity for every gene at differing levels of bootstrap support. Summary statistics for this gene tree–species tree similarity data were calculated in R v.3.1.2. We then isolated scores for gene found on the plasmid and the primary chromosome. This allowed us to compare gene tree–species tree variation by chromosome (i.e., plasmid V primary chromosome) using a Welch two-sampled t-test (test done in R v.3.1.2). This test was selected to determine if gene tree variation was significantly different than gene tree variation on the primary chromosome.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

We thank Marina Ascunce (University of Florida) and Katie Shepard (Shepard Institute) for providing lice used in this study and two anonymous reviewers for their comments on this manuscript. This work was supported by the National Science Foundation (DEB1310824 to B.M.B. & D.L.R.; DEB0845392 and DEB0717165 to D.L.R.; DEB0612938, DEB1239788, and DEB1342604 to K.P.J; and ABI-1458652 to T.W.). This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. XSEDE computational resources were provided through XSEDE grant ASC160042 to N.N.

References


