Effects of 16S rDNA sampling on estimates of endosymbiont lineages in sucking lice

Julie M Allen, J Gordon Burleigh, Jessica E Light, David L Reed

Co-evolution between insects and their endosymbiotic bacteria can be detected by constructing and comparing their phylogenetic trees. Even though taxon sampling can greatly affect phylogenetic and co-evolutionary inference, most hypotheses of endosymbiont relationships and estimates of the number of endosymbiont lineages within a host group have used only a small percentage of available bacterial sequences. Here we examined how different sampling strategies of *Gammaproteobacteria* sequences affect estimates of the number of endosymbiont lineages in parasitic sucking lice (Insecta: Phthirapatera: Anoplura). We estimated the number of louse endosymbiont lineages using both newly obtained and previously sequenced 16S rDNA bacterial sequences and more than 42,000 16S rDNA sequences from other *Gammaproteobacteria*. We also performed parametric and nonparametric bootstrapping experiments to examine the effects of phylogenetic error and uncertainty on these estimates. We found that sampling of 16S rDNA sequences affected the estimates of endosymbiont diversity in sucking lice until we reached a threshold of genetic diversity. Sampling by maximizing the diversity of 16S rDNA sequences was more efficient than simply randomly sampling available 16S rDNA sequences. Although simulation results support the finding of multiple endosymbiont lineages in sucking lice, the bootstrap results suggest that there is still uncertainty in estimates of the number of endosymbiont origins inferred from 16S rDNA alone.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1457v1 | CC-BY 4.0 Open Access | rec: 27 Oct 2015, publ: 27 Oct 2015

Effects of 16S rDNA sampling on estimates of endosymbiont lineages in sucking lice

Julie M. Allen[a,d*], J. Gordon Burleigh[b], Jessica E. Light[c], and David L. Reed[d]

[a] Illinois Natural History Survey, University of Illinois, 1816 South Oak St. Champaign, IL 61820, USA, juliema@illinois.edu

[b] Department of Biology, Box 118525, University of Florida, Gainesville, FL 32611, USA, gburleigh@ufl.edu

[c] Department of Wildlife and Fisheries Sciences, Texas A&M University, 210 Nagle Hall, 2258 TAMUS, College Station, Texas 77843, USA, jlight2@tamu.edu

[d] Florida Museum of Natural History, University of Florida, Museum Rd. at Newell Dr. Gainesville, Florida 32611, USA, dreed@flmnh.ufl.edu

[*]Corresponding author:

Julie M. Allen
Illinois Natural History Survey
University of Illinois
Champaign, IL 61820
Telephone: (217) 300-5651
Fax: (217) 265-4678
E-mail: juliema@illinois.edu.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1457v1 | CC-BY 4.0 Open Access | rec: 27 Oct 2015, publ: 27 Oct 2015

Abstract

Co-evolution between insects and their endosymbiotic bacteria can be detected by constructing and comparing their phylogenetic trees. Even though taxon sampling can greatly affect phylogenetic and co-evolutionary inference, most hypotheses of endosymbiont relationships and estimates of the number of endosymbiont lineages within a host group have used only a small percentage of available bacterial sequences. Here we examined how different sampling strategies of *Gammaproteobacteria* sequences affect estimates of the number of endosymbiont lineages in parasitic sucking lice (Insecta: Phthirapatera: Anoplura). We estimated the number of louse endosymbiont lineages using both newly obtained and previously sequenced 16S rDNA bacterial sequences and more than 42,000 16S rDNA sequences from other *Gammaproteobacteria*. We also performed parametric and nonparametric bootstrapping experiments to examine the effects of phylogenetic error and uncertainty on these estimates. We found that sampling of 16S rDNA sequences affected the estimates of endosymbiont diversity in sucking lice until we reached a threshold of genetic diversity. Sampling by maximizing the diversity of 16S rDNA sequences was more efficient than simply randomly sampling available 16S rDNA sequences. Although simulation results support the finding of multiple endosymbiont lineages in sucking lice, the bootstrap results suggest that there is still uncertainty in estimates of the number of endosymbiont origins inferred from 16S rDNA alone.

Key Words: Phylogenetics, 16S rDNA, Gammaproteobacteria, endosymbiont evolution, endosymbiosis, sucking lice, Anoplura

1. Introduction

Bacteria are among most common and diverse forms of life on the planet, and they are often found residing within, and interacting with, other organisms. For example, mutualisms between bacteria and insects are common, and in some cases bacteria inhabit specialized cells and provide a variety of benefits to their insect hosts (Buchner, 1965, Moran et al., 2008). The associations between bacteria and their hosts often occur over long evolutionary time scales, and phylogenetic trees of bacteria have helped scientists interpret the co-evolutionary history of these organisms with their hosts. In some cases, the phylogenetic tree of endosymbiotic bacteria corresponds with that of their insect hosts, suggesting co-speciation (Moran & Baumann, 1994). In other insect clades, the phylogenetic tree of the endosymbiont lineage does not match that of its host (Moran & Baumann, 1994; Lefevre et al., 2004; Hypsa & Krízek, 2007; Allen et al., 2009; Smith et al., 2013). When co-speciation is limited or absent, it is likely that there have been multiple origins of endosymbiosis within the bacteria. Thus, assessing the number of endosymbiont lineages in bacteria can provide new insights into the co-evolutionary history between insects and their internal bacteria.

Most of our understanding of the diversity of bacteria is based on a single locus, 16S rDNA (Schloss & Handelsman, 2004; Lozupone & Knight, 2007). In fact, many environmental studies identify bacterial diversity by sequencing a section of the 16S rDNA gene and comparing it to the enormous number of 16S rDNA sequences that reside in public databases. Although phylogenetic estimates using more genes, or even genomes, may provide a more complete phylogenetic perspective than a single gene analysis, the number of bacterial genomes sequenced represents far less phylogenetic breadth than the available 16S rDNA sequences (Klindworth et al., 2013). As more bacterial sequences become available, it is important to understand how taxonomic sampling may affect interpretations of bacterial-host associations.

Origins of bacterial endosymbiosis can be estimated by counting the independent endosymbiont clades in a bacterial phylogeny. Sampling of both endosymbiont and non-endosymbiont lineages in a phylogenetic analysis can greatly affect estimates of the number of endosymbiont origins either by the insertion of new sequences, which can break up or create clades of endosymbionts, or by directly affecting the resulting phylogenetic tree (e.g., Hillis, 1996; Hillis, 1998; Pollock et al., 2002; Zwickl & Hillis, 2002; Heath, Hedtke & Hillis, 2008). In this study, we focus on determining the number of distinct bacterial lineages (endosymbiont origins) found within parasitic sucking lice (Phthiraptera: Anoplura).

Sucking lice are wingless, blood-feeding insects that parasitize eutherian mammals. These lice have endosymbiotic bacteria that synthesize necessary amino acids and vitamins absent from the louse’s diet and are therefore thought to be required for louse survival. Previous studies have indicated that there are at least six different lineages of endosymbionts in sucking lice, all of which reside within *Gammaproteobacteria*, a class of gram-negative bacteria (Sasaki-Fukatsu et al., 2006; Allen et al., 2007; Hypsa & Kirizek, 2007; Allen et al., 2009; Fukatsu et al., 2009; Perotti et al., 2009). Phylogenetic studies show little concordance between the louse and bacteria trees (Hypsa & Kirizek, 2007; Allen et al., 2009); however, these studies estimated the number of louse endosymbiont lineages from only a tiny fraction (e.g., ~33-46 sequences) of the available 16S rDNA sequences.

Here we assembled a dataset with both new and previously studied 16S rDNA sequences from sucking louse endosymbionts and ~42,000 publicly available *Gammaproteobacteria* 16S rDNA sequences to determine the effect of sampling on our estimates of endosymbiont diversity. We counted the number of independent endosymbiont lineages on phylogenetic trees constructed from subsets of the entire sample of sequences. These subsets were created by either randomly sampling sequences or sampling sequences by maximizing genetic diversity. With these subsets, we also explored how different sequence selection strategies affect the estimates of endosymbiont origins. Lastly, we performed both parametric and nonparametric bootstrapping experiments to assess the role of error and uncertainty in estimates of the number of endosymbiont lineages.

2. Materials and Methods

2.1 Louse Endosymbiont Sampling and Sequencing

We obtained 23 louse specimens, representing 8 families and 21 species, from museums and mammal collectors (Table 1). The lice were washed three times in 500*u*l of 5% bleach and two times with sterile water to remove external bacteria (e.g., Meyer & Hoy, 2008). Lice were crushed and DNA extracted using a Qiagen micro kit (Cat No. 56304). We followed the manufacturer’s protocol except that the lice were placed in 80*u*l of Proteinase K (Qiagen) and incubated overnight on a heating block at 55°C, and the DNA was eluted in 50*u*l of sterile water. Water was used as a negative control for every extraction to ensure that there was no bacterial contamination. We amplified 16S rDNA from putative bacterial endosymbionts from each of the DNA samples using Stratagene Hi-Fidelity Master Mix (Cat No. 600650-51) with general bacterial primers 27F and

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1457v1 | CC-BY 4.0 Open Access | rec: 27 Oct 2015, publ: 27 Oct 2015

either 1525R or 1329R (Lane, 1991) at a final concentration of 0.7*u*M and total reaction volume of 50*u*l. Polymerase chain reaction (PCR) cycling conditions included an initial denaturation step at 95°C for two minutes followed by 40 cycles of denaturation at 95°C for 40 seconds, annealing at 50°C for 30 seconds, and extension at 72°C for two minutes, and a final extension step at 72°C for 30 minutes. The resulting PCR products were cloned using the Invitrogen Cloning Kit (Cat No. 45-0030), and 96 colonies per specimen were picked and sequenced at the University of Florida ICBR sequencing facility. The resulting 16S rDNA sequences were ~1,300 base pairs in length. All sequences will be submitted to GenBank (Accession numbers will be provided upon acceptance and sequences were uploaded with the manuscript).

We assessed if the 16S rDNA sequences amplified by PCR from the louse specimens came from endosymbionts based on their similarity to other endosymbiont sequences. If the most similar sequence from a BLAST search (Altschul et al., 1990) of the non-redundant nucleotide database in GenBank was from an endosymbiont, we identified the sequences as endosymbionts

We also downloaded from GenBank 12 endosymbiont sequences from sucking lice and Rhynchophthirina chewing lice (Accessions: DQ076661, DQ076662, DQ076665, DQ076664*,* EU827263, AB478979, EF110571, EF110573, DQ076667*,* DQ076666, EF110571, DQ076663; Hypsa & Kirizek 2007; Allen et al., 2009; Fukatsu et al., 2009). Rhynchophthirina is a suborder of blood-feeding chewing lice that is the sister group to Anoplura (Cruickshank et al., 2001; Barker et al., 2003; Johnson, Yoshizawa, & Smith, 2004; Yoshizawa & Johnson, 2010). Members of the suborder Rhynchophthirina parasitize eutherian mammals and are thought to have an endosymbiont serving a similar function as those in sucking lice.

2.2 16S rDNA Sampling and Alignments

We obtained an initial alignment of ~72,000 *Gammaproteobacteria* 16S rDNA sequences from the Ribosomal Database Project (Cole et al., 2005; http://rdp.cme.msu.edu/). We removed any sequences that contained fewer than 750 nucleotides and then deleted any columns in the alignment that contained fewer than 100 nucleotides. Next, we removed extra copies of identical sequences, so each remaining sequence was unique. We re-aligned the remaining sequences and the louse endosymbiont sequences using MUSCLE (Edgar, 2004). To do this, the sequences were split into 20 clusters of approximately equal size, and each cluster was aligned using the default settings in MUSCLE. The resulting alignments were edited manually. Profile alignments were then created using MUSCLE to combine the edited alignments. The resulting alignment of all sequences was checked by eye and edited manually. Regions of ambiguous alignment were removed, and any extra identical sequences were pruned from the alignment. This resulted in a final alignment of 42,626 sequences that was 1,476 characters in length. The final alignment is available in the Dryad repository (Data will be submitted to Dryad upon acceptance and was uploaded with the manuscript).

To determine how taxon sampling affects estimates of the number of endosymbiont lineages, we assembled five subsets of the 16S rDNA alignment. Our goal was to create taxonomic subsamples of increasing size such that each reflected the breadth of genetic diversity in the full alignment. To do this, we first clustered the sequences based on similarity using the QT-clustering algorithm (Heyer, 1999) implemented in RAxML-VI-

HPC version 7.0.4 (Stamatakis, 2006). We used five different thresholds for the sequence similarity clustering: 70%, 80%, 85%, 90%, and 95%. A higher threshold results in more clusters composed of more similar sequences. For each threshold, we sampled at least one sequence per cluster while ensuring that each subsample contained all louse endosymbiont sequences and all sequences included in the smaller clusters (e.g., the 85% cluster contained all sequences in the 80% cluster, the 80% cluster contained all sequences in the 70% cluster, etc.). In total, the sizes of the subsampled data sets were 39 taxa (70% cluster), 76 taxa (80% cluster), 217 taxa (85% cluster), 865 taxa (90% cluster) and 4,275 taxa (95% cluster). In order to compare this sampling strategy to a random taxon sampling strategy, we also created 100 data sets each of 76, 217, 865, and 4,275 taxa, each including all louse endosymbionts with the remaining sequences randomly selected from the full alignment.

2.3 Phylogenetic Analysis

We performed maximum likelihood (ML) phylogenetic analyses on each of the subsampled alignments using RAxML-VI-HPC version 7.0.4 with the GTRCAT nucleotide substitution model (Stamatakis, 2006). We also performed 200 non-parametric bootstrap replicates (Felsenstein, 1985) for each alignment using the same methods. For the ML search on the full data set (42,626 sequences), we used a parallelized version of RAxML for IBM BlueGene L clusters (Ott et al., 2007). This analysis took approximately 9 days to run on 256 processors at Iowa State University. A full bootstrap analysis was not feasible using this approach. Therefore, we created 100 nonparametric bootstrap data sets using HyPhy (Pond, Frost & Muse, 2005) and performed a ML analysis on these data sets using FastTree 2.1 with the GTRCAT model (Price, Dehal &

Arkin, 2010). The FastTree analyses used four minimum-evolution SPR rounds and the "-mlacc 2 –slownni" option to increase the search space of the NNI swaps in the ML analysis. Optimal trees from these analyses are available in the Dryad data repository (Data will be submitted upon acceptance and have been uploaded with the manuscript).

2.4 Number of Endosymbiont Lineages

For all ML and ML bootstrap trees, we inferred the number of independent endosymbiont clades with PAUP* (Swofford, 2003). Because endosymbionts do not spend any time outside of the insect and likely cannot transmit to new hosts (Moran et al., 2008), endosymbiosis was assumed to be a non-reversible binary character (i.e., non-endosymbionts can become endosymbionts, but endosymbionts cannot become non-endosymbionts). The placement of the root can affect the number of inferred origins of louse endosymbiosis, and the root of all sampled *Gammaproteobacteria* sequences is uncertain. Therefore, we calculated the number of louse endosymbiont origins using every possible rooting of the 16S rDNA tree. Re-rooting was done with a C++ program written for this analysis. Our estimate of the number of endosymbiont lineages is based on a root that implied the fewest louse endosymbiont origins.

The estimate of louse endosymbiont origins may change with increased taxonomic sampling due to either the insertion of new, non-endosymbiont sequences within an endosymbiont clade or changes in inferred relationships among endosymbionts. To help distinguish between these two possibilities, we took all optimal and bootstrap trees from the 80%, 85%, 90%, 95%, and full data sets and pruned them so that they would have the same taxon sampling as the smaller subsets. For example, the trees from 90% data set were pruned to create three data sets in which they would have only the taxa from 1) the

85% data set, 2) the 80%, and 3) the 70% data sets. Then we calculated the number of louse endosymbiont origins for each of the pruned trees. If the number of louse endosymbiont origins in the pruned trees equaled the number of endosymbiont origins estimated from the original data sets with the same taxa, then changes in the number of estimated endosymbiont origins in larger trees are caused by additional taxa breaking up endosymbiont clades. The taxon pruning was done with a Perl script and Newick utilities (Junier & Zdobnov, 2010).

2.5 Simulations

To evaluate if bias or error in our phylogenetic analyses could lead to erroneous estimates of louse endosymbiont origins, we used a parametric bootstrapping approach to estimate the number of louse endosymbiont origins we would expect to find if there was only one endosymbiont origin in lice. First, for the 70%, 80%, 85%, 90%, and 95% data sets, we performed a ML analysis in RAxML (Stamatakis, 2006) with all the louse endosymbiont sequences constrained to a single clade, which would imply a single origin of endosymbiosis. We then estimated the optimal branch length and GTR+I+G substitution model parameters for the 16S rDNA alignment used to infer the constraint tree using the resulting ML constraint topology for each data set and simulated 100 alignments of the same dimensions using HyPhy (Pond, Frost & Muse, 2005). On each simulated data set, we performed a ML analysis with RAxML and estimated the number of louse endosymbionts using the same protocol as we used on the empirical data. We then compared the number of endosymbiont origins inferred from our single-origin simulations to the number of origins inferred from the empirical data.

3. Results

3.1 Endosymbiont Sequences

We identified 18 endosymbiont sequences from 17 of the 23 louse specimens; two were found in a single louse (*Ancistroplax crocidurae*) and none were found in six specimens (Table 1). We used BLAST searches and AT content to assess if our newly acquired louse bacteria were from an endosymbiont (genomes of endosymbionts are often, but not always, AT-rich; Bentley & Parkhill, 2004; McCutcheon & Moran, 2012). All 18 sequences were most similar to other endosymbiont sequences in BLAST searches, and seven of these were most similar to other confirmed Anoplura endosymbionts (Table 1). All 18 sequences had ≥ 45% AT content, and 14 had ≥ 50% AT content, which is consistent with many endosymbionts (Moran & Baumann, 2000). We did not find any *Gammaproteobacteria* sequences that met our criterion in the louse genus *Hoplopleura*; however, we found *Alphaproteobacteria* sequences from the common louse pathogen *Bartonella* in four of the five *Hoplopleura* samples. Since our study was focused on *Gammaproteobacteria*, we did not use the *Alphaproteobacteria* sequences in our analyses. The 18 putative *Gammaproteobacteria* endosymbiont 16S rDNA sequences were combined with 12 sucking louse endosymbiont 16S rDNA sequences from GenBank so that all of the alignments used in the phylogenetic analyses contained 30 endosymbiont sequences from lice (Table 1).

3.2 Endosymbiont phylogeny

The phylogenetic relationships of the louse endosymbionts were largely consistent with previous studies. Our analysis revealed the same six lineages suggested in earlier publications, with similar topologies for these lineages. For example, endosymbionts from rat lice (*Polyplax* sp.) were nested within the genus *Legionella,* consistent with the findings of Hypsa and Krízek (2007). The endosymbiont lineage *Riesia* was monophyletic with a topology that suggests co-speciation with human, chimp, and gorilla lice (data not shown), similar to what was found in Allen et al. (2007). The 18 newly sequenced louse endosymbiont lineages revealed new clades of endosymbionts, all of which grouped close to *Arsenophonus* and other known insect endosymbionts including *Baumannia* and *Wigglesworthia* (the endosymbionts of sharpshooters and tsetse flies; Fig 1).

3.3 Estimates of Endosymbiont Lineages

The estimates of endosymbiont lineages increased from 2 in the ML trees from the 70% and 80% data sets with 39 and 76 taxa, respectively, to 10 in the ML trees of the 85%, 90%, 95%, and full data sets with 217, 865, 4,275 and 42,626 taxa, respectively (Fig. 2). For the randomly sampled data sets, the average number of louse endosymbiont lineages increased with the size of the data set up to 4,275 taxa. For the data sets with fewer than 1,000 sequences, the average estimates from the randomly sampled data sets were smaller than those found from the data sets of equal size that were sampled to maximize sequence diversity (Fig. 2). In the 4,275 taxon randomly sampled data set, the average number of endosymbiont lineages was similar to the estimate from the phylogenetically sampled data set with the same number of taxa (9.8 ± 1.3 SD and 10, respectively; Fig. 2).

Secondly, when the bootstrap replicates from the full data set were pruned to include only the sequences from the smaller data sets, the number of inferred endosymbiont lineages was similar to the original smaller size data sets (Fig. 3). These results suggest that as more sequences are added to the analyses, the numbers of endosymbiont lineages are changing because the new 16S rDNA sequences break up clades of endosymbionts, not because the new 16S rDNA sequences are changing the backbone topology of the tree.

Finally, examining the uncertainty in these estimates, we found notable variation in estimates of endosymbiont lineages across the bootstrap replicates (Fig. 2). The standard deviation of number of lineages among bootstrap replicates was lowest for the 70% dataset, highest for the 80% dataset. The parametric bootstrapping analysis of the data sets simulated from a tree with a single origin of endosymbiosis in lice resulted in estimates of the number of endosymbiont lineages ranging from one (for the smaller datasets) to at most three endosymbiont lineages for the larger data sets (Fig. 2). This result indicates that if there was only a single origin of endosymbiosis, we would expect our estimates from 16S rDNA would reflect at most only a few origins. Because the empirical estimates of endosymbiont origins far exceed the estimates in the single-origin simulation, it is likely that there were multiple origins of endosymbiosis in sucking lice.

4. Discussion

Phylogenetic trees of bacteria have helped reveal the origins of symbioses and the co-evolutionary history between these organisms and their hosts. While the abundance of 16S rDNA sequences enables us to build enormous phylogenetic trees of bacteria, few studies have explored how sampling of available 16S rDNA sequences affects our interpretations of the co-evolutionary history of bacteria and their hosts. Taxon sampling can greatly affect phylogenetic reconstruction (Hillis, 1996; Hillis 1998; Pollock et al., 2002; Zwickl & Hillis, 2002; Heath, Hedtke & Hillis, 2008). Furthermore, new bacterial sequences can reveal new clades of endosymbionts, either by adding in new endosymbiont lineages or adding in non-endosymbionts to break up apparent endosymbiont clades. Therefore, it is important to explore how taxon sampling affects our estimates of endosymbiont lineages.

Overall, our estimates of endosymbiont lineages remain relatively unchanged as long as the tree contains a minimal level of genetic diversity of *Gammaproteobacteria*. For example, once we sampled ~200 sequences by maximizing sequence diversity, adding additional sequences had little effect on our estimates of the numbers of louse endosymbiont lineages (10; Fig.2). In contrast, if we added randomly selected sequences, we needed to sample at least ~4,000 sequences before the estimates of endosymbiont lineages converged to 10 (Fig. 2). This result emphasizes the importance of addressing the question of number of independent endosymbiont origins in the context of all *Gammaproteobacteria* sequence diversity. If the 16S rDNA sequences are chosen to maximize their diversity, fewer sequences may be needed to infer the number of endosymbiont lineages.

16S rDNA is the barcoding gene used to identify unique bacterial lineages, and much of our understanding of bacterial diversity comes from this gene (Klindworth et al., 2013). Therefore, it is uniquely useful for estimating the total number of endosymbiont lineages among *Gammaproteobacteria*. However, phylogenetic trees with thousands of leaves constructed from a single locus likely include much error and uncertainty. It is unclear how much this topological error or uncertainty affects estimates of the number of endosymbiont lineages. We addressed this question using nonparametric and parametric bootstrapping experiments. First, we calculated the number of implied louse endosymbiont origins on all bootstrap trees to assess how topological uncertainty might affect the analyses. Although the estimates of endosymbiont lineages varied among bootstrap replicates (Fig. 2), no bootstrap replicate in the data sets with more than 865 taxa implied fewer than 8 louse endosymbiont lineages. In other replicates, the number of estimated endosymbiont lineages exceeded 15, suggesting that error can inflate estimates of endosymbiont origins (Fig. 2).

We also performed a parametric bootstrapping experiment to assess the number of endosymbiont origins we would infer if there were only a single origin in sucking lice. In some cases, analyses of the simulated data sets inferred more than a single origin, but they never inferred more than three origins on any simulated data set (Fig 2). This suggests that error in the topology will not necessarily radically affect estimates of the number of origins of endosymbiosis.

Our work demonstrates that the number of inferred endosymbiont lineages may be accurate if the diversity of sequence sampling is sufficient. Still, it is unclear how many more endosymbiont lineages we would find with greater sampling. Adding any single new sequence from other *Gammaproteobacteria* could reveal additional endosymbiont origins. Furthermore, not all families of Anoplura have been sequenced, and additional sampling may reveal more endosymbiont lineages. Even still, our estimate of at least 10 endosymbiont origins is large compared to other insect/endosymbiont assemblages with one or only a few endosymbiont lineages (e.g. aphids and *Buchnera*; Moran & Baumann 1994). Although it is impossible to determine with certainty the nature of the relationship of the bacteria with the host (e.g., mutualistic primary endosymbionts or facultative) from only 16S rDNA sequences, our results suggest the possibility that co-evolution of bacterial endosymbionts in sucking lice is an extremely labile process.

While 16S rDNA is, and will likely be for the foreseeable future, the most widely sequenced gene for bacterial identification, additional genes and even genomic sequencing will enable phylogenetic estimates of the bacteria based on many loci. Although these data may ameliorate biases or error associated with 16S rDNA and reduce uncertainty in phylogenetic estimates, they will unlikely rival the diversity found in 16S rDNA, which may be critical for estimating the number of endosymbiont lineages. In the future, combining the sampling of 16S rDNA with the phylogenetic power of large genomic data will likely provide a more complete picture of the evolutionary history of insect associated bacteria.

Acknowledgements

We thank L. Durden, L. Barrow, S. Bush, K. Shepherd, F. Leendertz, R. Ream, C. Chapman, T. Manning, D. Hafner, J. Hafner, C. Worman, and J. Skinner for obtaining samples of sucking lice. We thank E. Bess, B. Boyd, K. Han, S. Hilber, K. Krisko, C. Rittschof, K. Olson, J. Oswald, C. Phillips, J. Pino, J. Saunders, A. Soto-Centeno and N. Wang for help with the alignments. Avinash Ramu wrote the program to re-root trees and B. Boyd provided helpful comments on the manuscript. For computational support Andre Wehe and the Iowa State University BlueGene cluster, the UFGI Fisher cluster, and statistical support Terry Allen at Weber State University.

Funding Contributions:

This work was funded in part by grants from the National Science Foundation to DLR (DEB 0717165 and DEB 0845392) and JEL (DEB 0308878 and DEB 0717165).

References

Allen, J.M., Light, J.E., Perotti, M.A., Braig, H.R., Reed, D.L. 2009. Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. PLoS One 4:e4969.

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. Microb. 73:1659-64.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Barker, S.C., Whiting, M., Johnson, K.P., Murrell, A. 2003. Phylogeny of the lice (Insecta, Phthiraptera) inferred from small subunit rRNA. Zool. Scr. 32:407- 414.

Bentley, S.D., Parkhill, J. 2004. Comparative genomic structure of prokaryotes. Ann. Rev. Gen. 38:771-91.

Buchner, P. 1965. Endosymbiosis of animals with plant microorganisms. Intersci. Publ. Inc., New York.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., Garrity, G.M., Tiedje, J.M. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 33:D294-6.

Cruickshank, R.H., Johnson, K.P., Smith, V.S., Adams, R.J., Clayton, D.H., Page, R.D. M. 2001. Phylogenetic Analysis of Partial Sequences of Elongation Factor 1 α Identifies Major Groups of Lice (Insecta: Phthiraptera). Mol. Phylogenet. Evol. 19:202-215.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-7.

Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39:783-791.

Fukatsu, T., Hosokawa, T., Koga, R., Nikoh, N., Kato, T., Hayama, S., Takefushi, H., Tanaka, I. 2009. Intestinal endocellular symbiotic bacterium of the macaque louse *Pedicinus obtusus*: Distinct endosymbiont origins in anthropoid primate lice and the old world monkey louse. Appl. Environ. Microbiol. 75:3796-9.

Heath, T.A., Hedtke, S.M., and Hillis, DM. 2008. Taxon sampling and the accuaracy of phylogenetic analysis. J. of System and Evol. 46(3):239-257.

Heyer, L.J. 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9:1106-1115.

Hills, D.M. Inferring complex phylogenies. 1996. Nature. 383:130-131.

Hillis D.M. 1998. Taxonomic sampling, phylogenetic accuracy and investigator bias. Syst. Biol. 74:3-8.

Hypsa, V., Krízek, J. 2007. Molecular evidence for polyphyletic origin of the primary symbionts of sucking lice (Phthiraptera, Anoplura). Microb. Ecol. 54:242-51.

Johnson, K.P., Yoshizawa, K., Smith, V.S. 2004. Multiple origins of parasitism in lice. Proc. R. Soc. Lond. B. 271:1771-1776.

Junier, T., Zdobnov, E.M. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26:1669-70.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F.O. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nuc. Acid. Res. 41:e1. doi:10.1093/nar/gks808.

Lane, D.J. 1991. 16S/23S rRNA sequencing. Pp. 115-175 *in* E. Stackebrandt and M. Goodfellow, eds. Nucleic acid techniques in bacterial systematics. John Wiley Sons, New York.

Lefèvre, C., Charles, H., Vallier, A., Delobel, B., Farrell, B., A. Heddi, A.. 2004. Endosymbiont phylogenesis in the Dryophthoridae weevils: evidence for bacterial replacement. Mol. Biol. Evol. 21:965-73.

Lozupone, C.A., Knight R. 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. USA 104:11436-40.

McCutcheon, J.P., Moran, N.A. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13-26.

Meyer, J.M., Hoy, M.A. 2008. Removal of fungal contaminants and their DNA from the surface of *Diaphorina citri* (hemiptera: psyllidae) prior to a molecular survey of endosymbionts. Fla. Entomol. 91:702-705.

Moran, N.A., Baumann, P. 1994. Phylogenetics of cytoplasmically inherited microorganisms of arthropods. Trends Ecol. Evol. 9:15-20.

Moran, N.A., Baumann, P. 2000. Bacterial endosymbionts in animals. Curr. Opin. Microbiol. 3:270-5.

Moran, N.A., McCutcheon, J.P., Nakabachi, A. 2008. Genomics and evolution of heritable bacterial symbionts. Ann. Rev. Gen. 42:165-90.

Ott, M., Zola, J., Aluru, S., Stamatakis, A. 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. Page 1 Proceedings of the 2007 ACM/IEEE conference on Supercomputing. ACM Press, Reno, Nevada.

Perotti M.A., Kirkness, E.F., Reed, D.L., Braig, H.R. 2009. Endosymbionts of lice. *In* Bourtzis, K., T.A. Miller, eds. *Insect Symbiosis,* 3:205-219. Boca Raton, FL: CRC Press

Pond, S.L.K., Frost, S.D.W., Muse, S.V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-9.

Price, M.N., Dehal, P.S., Arkin, A.P. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Pollock, D.D., Zwickl D.J., McGuie J.A. and Hillis D.M. 2002. Increased Taxon sampling is advantageors for phylogenetic inverece. Syst. Biol. 51(4):664-671.

Sasaki-Fukatsu K, Koga R, Nikoh N, Yoshizawa K, Kasai S, Mihara M, et al. Symbiotic bacteria associated with stomach discs of human lice. Appl Environ Microbiol. 2006;72:7349–7352.

Schloss, P.D., Handelsman, J. 2004. Status of the microbial census. Microbiol. Molec. Biol. Rev. 68:686-691.

Smith, W.A., Oakeson, K.F., Johnson, K.P., Reed, D.L., Carter, T., Smith, K.L., Koga, R., Fukatsu, T., Clayton, D.H., Dale, C. 2013. Phylogenetic Analysis of Symbionts in Feather-Feeding Lice of the Genus Columbicola: Evidence for Repeated Symbiont Replacements. BMC Evol. Biol.13:109.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-90.

Swofford, D.L. 2003. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer, Sunderland, Massachusetts.

Yoshizawa, K., Johnson, K.P. 2010. How stable is the “Polyphyly of Lice” hypothesis (Insecta: Psocodea)?: a comparison of phylogenetic signal in multiple genes. Molec. Phylogenet. Evol. 55:939-51.

Zwickl, D.J. and Hillis, DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51(4):588-98.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1457v1 | CC-BY 4.0 Open Access | rec: 27 Oct 2015, publ: 27 Oct 2015

Table 1(on next page)

Table of Anoplura endosymbiont sequences

Table 1. Family and species of sucking lice (Phthiraptera: Anoplura) from which endosymbionts were targeted. Also indicated are the collection locality, louse taxon label (for use in the laboratory), mammalian host, presence of putative endosymbiont (where the superscript "B" indicates that *Bartonella*, a louse pathogen, was sequenced), percent AT content, if the top hit from a BLAST search was an endosymbiont, and finally if the top hit from the BLAST search was an endosymbiont from a sucking louse.

Louse Family and Species Country and State	Taxon Label	Host (Order: Family) Museum Voucher [if known])	Endo Present	%AT	BLAST Endo	BLAST Anoplura
Echinophthiriidae						
Proechinophthirus fluctus (USA: AK)	Echin3.17.09.2	*Callorhinus ursinus* (Carnivora: Otariidae)	Yes	45%	Yes	No*
Haematopinidae						
Haematopinus suis (USA: FL)	Hpsu7.14.09.4	*Sus scrofa* (Artiodactyla: Suidae)	Yes	52%	Yes	Yes*
Hoplopleuridae						
Ancistroplax crociduræ 1 (Vietnam)	Axcro4.26.09.1	*Crocidura* sp. (Soricomorpha: Soricidae)	Yes	50%	Yes	No
Ancistroplax crociduræ 2 (China)	Axsp7.14.09.5	*Crocidura attenuata* (Soricomorpha: Soricidae)	Yes (2)	49%, 45%	Yes,Yes	No,No
Hoplopleura ferrisi 2 (MX: Puebla)	Hofer7.14.09.8	*Peromyscus difficilis* (Rodentia: Cricetidae; LSUMZ 36247)	No	-		
Hoplopleura hirsuta (USA: TX)	Hosp4.17.09.7	*Sigmodon hispidus* (Rodentia: Cricetidae; LSUMZ 36377)	No	-		
Hoplopleura onychomydis (USA: AZ)	Hoony8.27.08.6	*Onychomys torridus* (Rodentia: Cricetidae; NMMNH 4394)	No	-		
Hoplopleura reithrodontomydis 2 (USA: AZ)	Hosp7.14.09.6	*Reithrodontomys* sp. (Rodentia: Cricetidae; NMMNH 4411)	No	-		
Hoplopleura sicata (China)	Hosic7.14.09.9	*Niviventer fulvescens* (Rodentia: Muridae)	No	-		
Linognathidae						
Linognathus spicatus (Zimbabwe)	Linog6.22.09.1	*Connochaetes taurinus* (Artiodactyla: Bovidae)	Yes	52%	Yes	No*
Pedicinidae						
Pedicinus pictus 1 (Ivory Coast)	Qnpic3.31.08.1	*Piliocolobus badius* (Primates: Cercopithecidae)	Yes	54%	Yes	Yes*
Pedicinus pictus 2 (Ivory Coast)	Qnpic6.30.09.2	*Colobus polykomos* (Primates: Cercopithecidae)	Yes	53%	Yes	Yes
Pedicinus pictus 3 (Ivory Coast)	Qnsp3.31.08.3	*Colobus polykomos* (Primates: Cercopithecidae)	Yes	54%	Yes	Yes
Pediculidae						
Pediculus humanus capitis (USA: FL)	Pdcap9.20.05.2NW	*Homo sapiens* (Primates: Hominidae)	Yes	51%	Yes	Yes
Pediculus humanus humanus (USA: MD)	Pdhum5.19.05.2	*Homo sapiens* (Primates: Hominidae)	Yes	51%	Yes	Yes
Polyplacidae						
Fahrenholzia ehrlichi 1 (USA: TX)	Fzehr8.20.08.1	*Liomys irroratus* (Rodentia: Heteromyidae; LSUMZ 36395)	Yes	52%	Yes	No
Fahrenholzia ehrlichi 2 (MX:Puebla)	Fzehr6.30.09.4	*Liomys irroratus* (Rodentia: Heteromyidae; LSUMZ 36299)	Yes	51%	Yes	No

Table 1 Continued:

Louse Family and Species	Taxon	Host (Order: Family)	Endo	%AT	BLAST	BLAST
Country and State	Label	Museum Voucher [if known])	Present		Endo	Anoplura
Linognathoides marmotae 1 (USA: CO)	Lnlae6.30.09.3	*Marmota flaviventris* (Rodentia: Sciuridae)	Yes	54%	Yes	No
Lemurpediculus verruculosus 1 (Madagascar)	Lesp4.26.09.2	*Microcebus rufus* (Primates: Cheirogaleidae)	Yes	53%	Yes	No
Neohaematopinus sciuropteri (USA: OR)	Nescp6.30.09.5	*Glaucomys sabrinus* (Rodentia: Sciuridae)	Yes	53%	Yes	No
Neohaematopinus neotomae (USA: CA)	Neneo8.20.08.2	*Neotoma lepida* (Rodentia: Cricetidae; MLZ 1921)	No	-		No
Sathrax durus (Vietnam)	Sathrax4.26.09.3	*Tupaia belangeri* (Scandetia: Tupaiidae)	Yes	45%	Yes	No
Pthiridae						
Pthirus gorillae (Uganda)	Ptgor9.14.08.1	*Gorilla gorilla* (Primates: Hominidae)	Yes	53%	Yes	Yes

USA=United States (AK=Alaska, AZ=Arizona,CA=California, CO=Colorado,FL=Florida, MD=Maryland, OR=Oregon, TX=Texas); MX=Mexico

MLZ = Moore Laboratory of Zoology

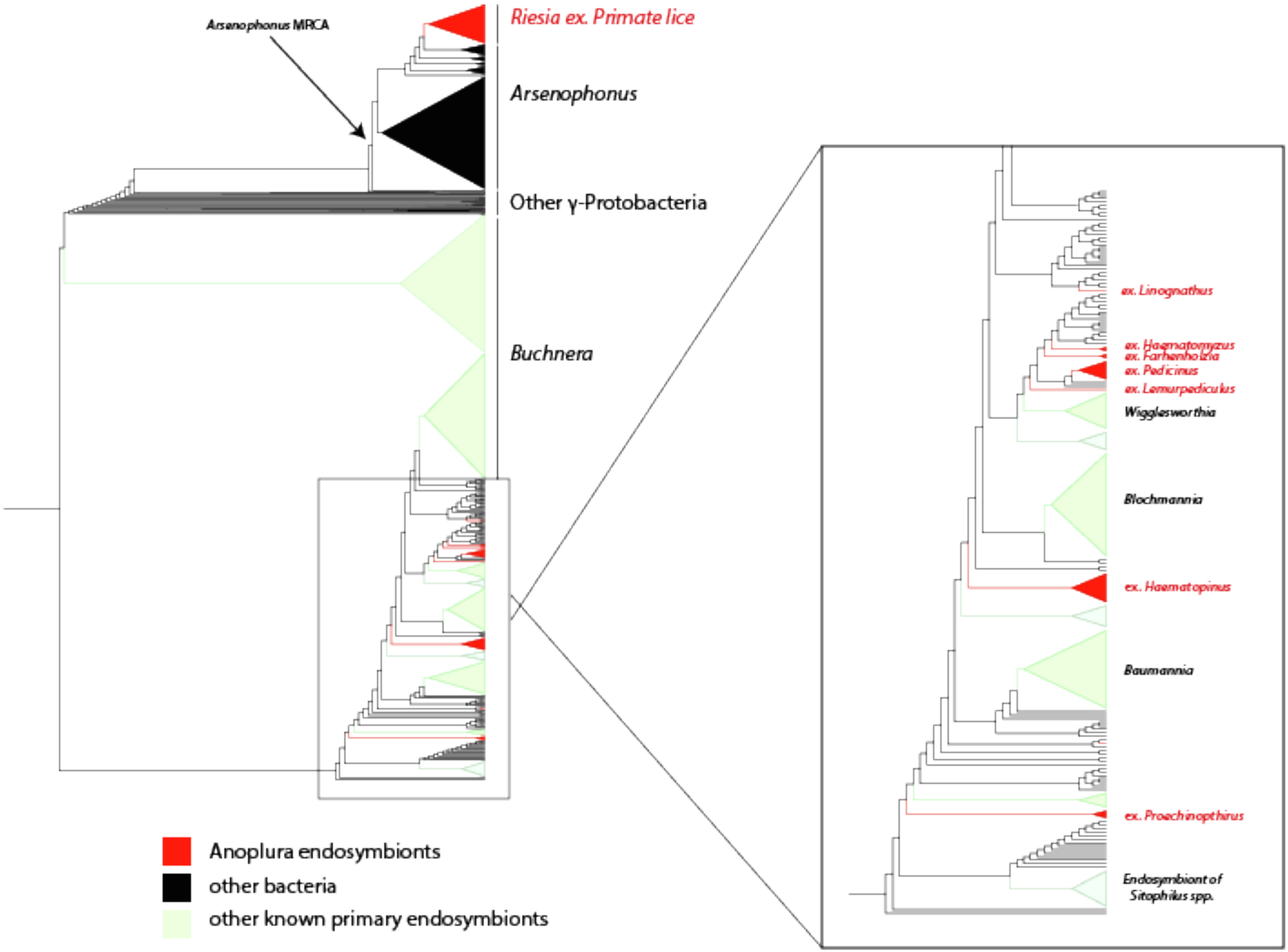
LSUMNZ=Louisiana State University Museum of Zoology

NMMNH=New Mexico Museum of Natural History

Figure 1(on next page)

Subset of large phylogenetic tree showing placement and close relatives of endosymbiotic bacteria in Anoplura

Figure 1: A subtree of the full 42,266 *Gammaproteobacteria* tree showing 9 of the 10 endosymbiont lineages from sucking lice (red). For all louse endosymbionts, the louse host genus or group is indicted. All of these sequences cluster together either within or near other known endosymbiont lineages (green) and *Arsenophonus*, a clade of insect bacterial endosymbionts; the arrow points to the Most Recent Common Ancestor (MRCA) of this clade. The 10th lineage of endosymbiont clusters with the genus *Legionella,* which is not shown due to space constraints.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1457v1 | CC-BY 4.0 Open Access | rec: 27 Oct 2015, publ: 27 Oct 2015

Figure 2(on next page)

Box Plots showing number of endoysmbiont lineages in differently sampled datasets

Figure 2. The number of sucking louse endosymbiont lineages inferred from phylogenetic trees with different sampling. The number of taxa in each alignment is plotted on a $\log_{10}$ scale. Boxplots represent the number of endosymbionts calculated from either the 200 bootstrap replicates for the phylogenetically sampled data sets (in black), across the 100 randomly sampled data sets (red) or the simulated data sets (blue). Boxes represent 50% of the data; whiskers extend to 1.5 times the interquartile range representing 95% of the data, and * shows the average. “X” corresponds to the number of lineages calculated from the ML tree for each data set.

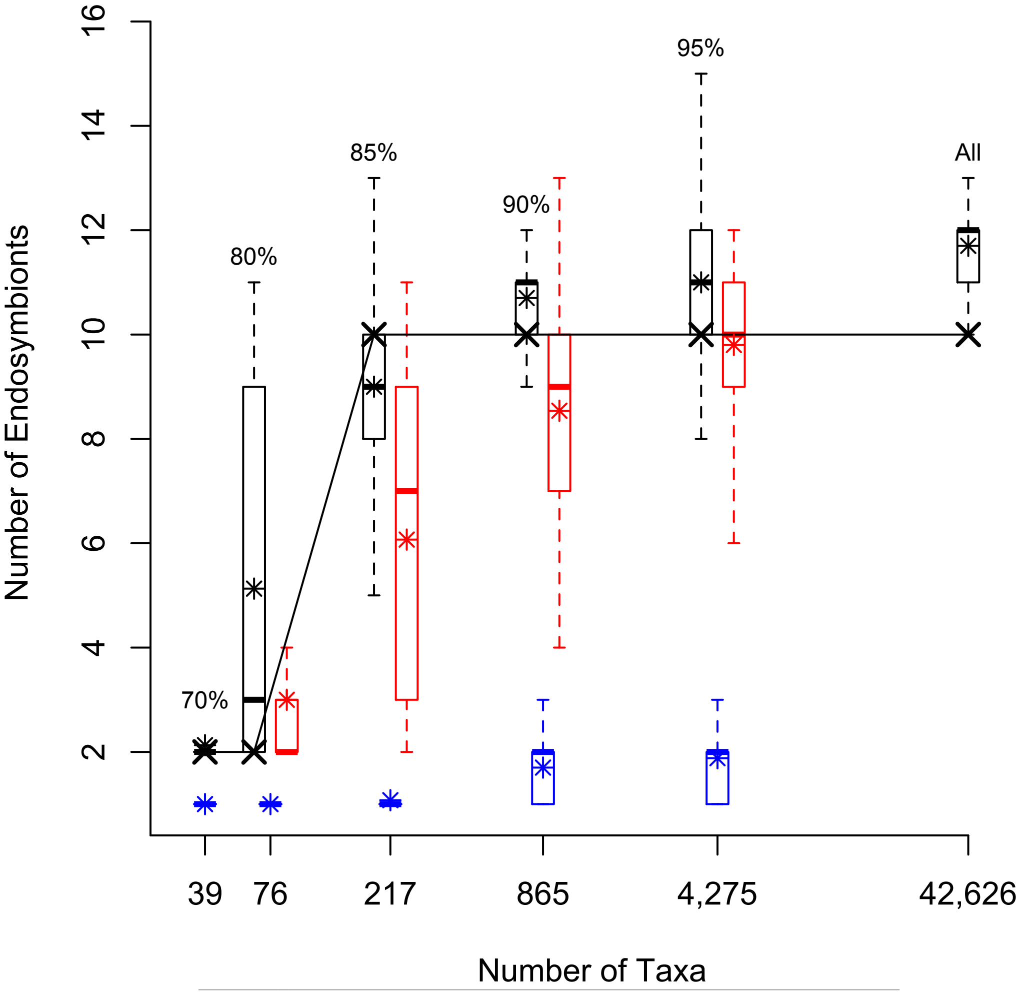
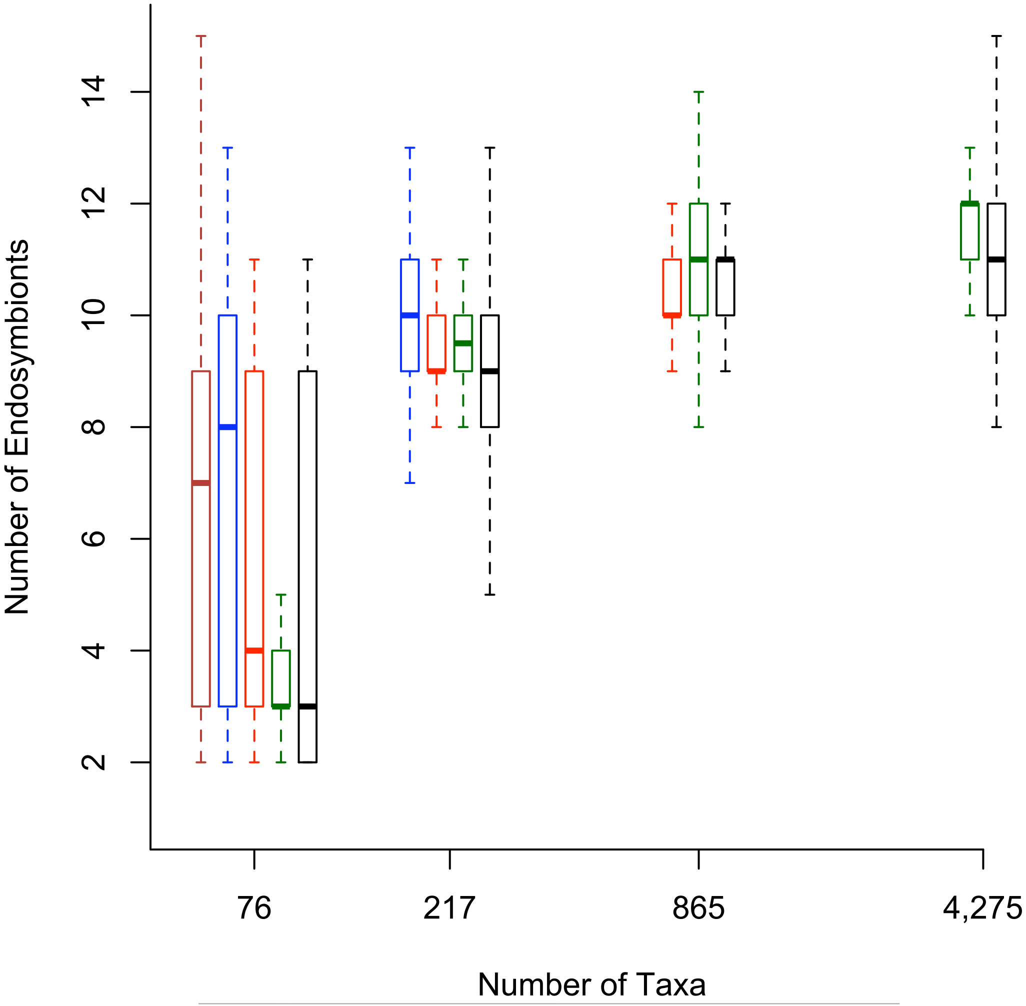

Number of Endosymbionts
16
14
12
10
8
6
4
2
70%
80%
85%
90%
95%
All
39
76
217
865
4,275
42,626
Number of Taxa

Figure 3(on next page)

Box plots with number of lineages for reduced phylogenetic trees

Figure 3. The number of sucking louse endosymbiont lineages found for reduced phylogenetic trees. Boxplots represent the number of endosymbiont lineages calculated from 200 bootstrap replicates for the data sets. The 200 bootstrap trees for each data set were then pruned to the taxa found in the smaller data sets and the number of endosymbiont lineages counted. The original data sets are plotted in black. The reduced full data sets are in green, reduced 95% data sets are in red, reduced 90% data sets in blue, and reduced 85% data sets in brown. Boxes represent 50% of the data and whiskers extend to 1.5 times the interquartile range, representing 95% of the data.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1457v1 | CC-BY 4.0 Open Access | rec: 27 Oct 2015, publ: 27 Oct 2015