Phylogenomics from Whole Genome Sequences Using aTRAM

Julie M. Allen[1*], Bret Boyd[1,2], Nam-phuong Nguyen[3], Pranjal Vachaspati[4], Tandy Warnow[3,4,10], Daisie I. Huang[5], Patrick G.S. Grady[1], Kayce C. Bell[6], Quentin C.B. Cronk[5], Lawrence Mugisha[7, 8], Barry R. Pittendrigh[9], M. Soledad Leonardi[10], David L. Reed[2], and Kevin P. Johnson[1]

1. *Illinois Natural History Survey, University of Illinois at Urbana-Champaign, Urbana IL*
2. *Florida Museum of Natural History, University of Florida, Gainesville, FL*
3. *Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL*
4. *Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL*
5. *Biodiversity Research Centre, University of British Columbia, Vancouver, Canada*
6. *Department of Biology and Museum of Southwestern Biology, University of New Mexico, Albuquerque, NM 87131, USA*
7. *Conservation & Ecosystem Health Alliance (CEHA), Kampala, Uganda*
8. *College of Veterinary Medicine, Animal Resources & Biosecurity (COVAB), Makerere University, Uganda.*
9. *Department of Entomology Michigan State University, East Lansing MI 48823, United States*
10. *Instituto de Biología de Organismos Marinos, Centro Nacional Patagónico, Puerto Madryn, Argentina*
11. *Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL*

* Corresponding Author: Julie M Allen; *Illinois Natural History Survey, University of Illinois at Urbana-Champaign, Urbana IL;* juliema@illinois.edu (352)-359-7655.

ABSTRACT

Novel sequencing technologies are rapidly expanding the size of datasets that can be applied to phylogenetic studies. Currently the most commonly used phylogenomic approaches involve some form of genome reduction. While these approaches make assembling phylogenomic datasets more economical for organisms with large genomes, they reduce the genomic coverage and thereby the long-term utility of the data. Currently, for organisms with moderate to small genomes (<1000 Mbp) it is feasible to sequence the entire genome at modest coverage (10-30X). Computational challenges for handling these large datasets can be alleviated by assembling targeted reads, rather than assembling the entire genome, to produce a phylogenomic data matrix.

Here we demonstrate the use of automated Target Restricted Assembly Method (aTRAM) to assemble 1,107 single copy ortholog genes from whole genome sequencing of sucking lice (Anoplura) and outgroups. We developed a pipeline to extract exon sequences from the aTRAM assemblies by annotating them with respect to the original target protein. We aligned these protein sequences with the inferred amino acids and then performed phylogenetic analyses on both the concatenated matrix of genes and on each gene separately in a coalescent analysis. Finally, we tested the limits of successful assembly in aTRAM by assembling 100 genes from close to distantly related taxa at high to low levels of coverage.

Both the concatenated analysis and the coalescent-based analysis produced the same tree topology, which was consistent with previously published results and resolved weakly supported nodes. These results demonstrate that this approach is successful at developing phylogenomic datasets from raw genome sequencing reads. Further, we found that with coverages above 5 – 10X, aTRAM was successful at assembling 80 – 90% of the contigs for both close and distantly related taxa. As sequencing costs continue to decline, we expect full genome sequencing will

become more feasible for a wider array of organisms, and aTRAM will enable mining of these genomic datasets for an extensive variety of applications, including phylogenomics.

KEY WORDS: phylogenomics, genome sequencing, aTRAM, gene assembly

The advent of next-generation sequencing (NGS) has facilitated rapid generation of genomic datasets to address difficult phylogenetic problems (Jarvis et al. 2014; Misof et al. 2014; Prum et al. 2015). However, there is considerable variation in not only the type of data collected, but also how it is curated and analyzed. For example, one common practice is to sequence only specific genetic markers. In these 'genome reduction' approaches, only targeted markers are sequenced, reducing the overall amount of sequencing data needed and therefore the cost. These approaches also reduce the complexity of the genomic dataset, simplifying downstream assembly and analysis. One frequently used method of genome reduction involves sequencing transcribed mRNAs (the transcriptome; Morozova et al. 2009), which has yielded hundreds of gene sequences for deep scale phylogenomic analyses (e.g. Misof et al. 2014). Other genome reduction approaches, such as ultra-conserved elements (UCEs; Faircloth et al. 2012, 2015) and anchored hybrid enrichment (AHE; Lemmon et al. 2012) target particular regions of the genome to enrich and reduce the fraction that is sequenced. These approaches have proved useful in collecting data for difficult phylogenetic problems, particularly in organisms with larger genomes (Misof et al. 2014; Prum et al. 2015).

While able to produce a dramatic increase in the number of phylogenetically informative markers, genome reduction techniques have some limitations. For transcriptome sequencing, a large quantity of high quality RNA is needed. This quantity is often not feasible for small organisms (e.g. small insects) where multiple individuals need to be pooled to obtain sufficient quantities of mRNA. If there is unrecognized (cryptic) genetic variation or even the presence of multiple species, pooling multiple individuals will create mixed samples. In addition, mRNA is not stable in samples stored using more traditional collection techniques, including ethanol preserved or pinned specimens, making it difficult or impossible to include rare or extinct

species on the basis of transcriptomes. DNA enrichment techniques, such as UCEs or AHE, have an advantage over transcriptome sequencing in that degraded and smaller samples can be used. While approaches using whole genome amplification may ameliorate this problem, they often produce chimeric sequences and amplification biases (Laskin and Stockwell 2007; Zhang et al. 2014). Furthermore, these approaches rely on probe design, which can be difficult when the group of interest has few genomic resources from which to design probes. One final drawback of all genome reduction techniques is the reduced dataset produced. The data is often generated with the sole purpose of phylogenetic estimation. Therefore, future studies wishing to add taxa to the phylogeny must target the same loci, similar to limitations of targeted Sanger sequencing projects. In addition, the loci targeted may have limited utility outside of phylogenetics.

Ultimately, having unbiased representation of genomic DNA (gDNA) prior to sequencing has major advantages over genome reduction methods. For example, with whole genome sequencing there is no need for marker development and optimization, reducing the amount of time and cost prior to sequencing. Further, all taxa can be sequenced using the same approach not simply those for which the probe design has been optimized, or those for which samples were collected appropriately for RNA. In addition, applications beyond phylogenomics can be pursued using these same genome sequences (e.g. study of genome organization, study of non-targeted genes, assembly of mitochondrial and microbial associates), allowing these datasets to be repeatedly mined for a wide variety of informative biological data. If sequencing costs continue to decline according to current trends (Wetterstrand 2013), it will eventually be feasible to sequence the full genome of all organisms regardless of genome size, providing vast amount of sequences from which to assemble multiple datasets. For species with small genomes (<1000 Mbp) it is already economical to sequence the entire genome with enough coverage to assemble

different datasets (Allen et al. 2015; Boyd et al., 2014, 2016). With these organisms, the datasets can also be used for many future studies, having no limitations on the types of loci from gDNA available to assemble.

While whole genome sequencing has fewer limitations on the types of data produced, downstream computational processing of these larger datasets can be limiting. Several studies have estimated phylogenomic trees by mining data from draft or complete genome assemblies (e.g. Jarvis et al. 2014; Niehuis et al. 2012). Unfortunately, the difficulties and time associated with genome assemblies can be prohibitive. Currently, full genomes can be assembled either *de novo* or by aligning the short reads to a closely related genome. While reference-based assemblies are faster than *de novo* approaches, they can only be accomplished with the genome of a closely related taxon. Further, if there are insertions, inversions, or other genome arrangements in the newly sequenced genome, those may not be assembled or may be mis-assembled, as only reads that map to the reference are considered. Therefore structural variation in the newly sequenced genome may not be detected (Pop 2009). While the issues with reference-based assemblies may be ameliorated with *de novo* assembly, this process is not only time consuming but also computationally difficult, and errors can be introduced while trying to reconstruct an entire genome (Salzberg and Yorke 2005; Alkan et al. 2011). In fact, to highlight the difficulty of this process, it has been suggested that different *de novo* assemblies be constructed with alternate settings to take into account uncertainty in the genome assembly process (Howison et al. 2013; 2014). Fortunately, for many phylogenomic studies a fully assembled genome is not necessary, rather only phylogenetically informative loci are needed. Therefore, the genome assembly process can potentially be bypassed along with any errors that may be introduced during the assembly and the typically large investment of time required.

Therefore, methods to assemble specific loci of interest from whole genome sequencing reads will be valuable for quickly assembling different datasets.

We developed the *automated Target Restricted Assembly Method* (aTRAM) to assemble small-targeted regions of a short-read genome dataset (Johnson et al. 2013; Allen et al. 2015). By targeting and assembling only reads that match specific loci, the assembly process is faster and less complicated. To produce these localized assemblies, aTRAM relies on BLAST searches (Altschul et al. 1990) to identify reads that contain potentially homologous bases to a reference sequence. The reads are then assembled, *de novo*, into a contiguous sequence (contig). The resulting contig replaces the original reference and the process is repeated multiple times. This iterative BLAST-assembly method allows for quick assembly of small genomic regions. Additionally, a closely related reference sequence is not necessary, because protein blast (TBLASTN) is successful for highly divergent taxa, allowing protein-coding gene assemblies across distantly related taxonomic groups (Allen et al. 2015). Further, because the reads are assembled *de novo*, the assemblies are not as tightly restricted to the reference sequence. Therefore, inversions or other structural differences in the newly sequenced genome will be revealed.

Here we describe the implementation of aTRAM for generating a phylogenomic dataset of 1:1 single-copy protein-coding orthologous genes. We use this dataset to illustrate the potential of aTRAM to generate a phylogenomic dataset for hundreds to thousands of genes in a relatively cost effective and timely manner and generate a phylogeny from these data. We also assess the limitations of aTRAM for assembling genes with varying levels of sequencing depth and divergence from the reference. Therefore, in addition to generating a phylogenomic dataset and phylogenetic analysis, we also provide an assessment of the effect that genetic divergence

and genome coverage has on aTRAM assembly success. These results provide basic guidelines for researchers wanting to use this software on various target taxa.

Focal Group

We constructed a dataset of raw next-generation short read sequences from 15 species of lice (Insecta: Phthiraptera), including 13 from mammal sucking lice (Anoplura) and two bird chewing lice (Ischnocera, outgroup). These sucking lice are obligate ectoparasites of mammals and feed on blood by piercing the skin (Snodgrass 1944). The body size of these lice is extremely small (<2 mm length), so obtaining a large quantity of good quality specimens from any given species is problematic, making RNA (transcriptome) sequencing infeasible. Furthermore, lice evolve relatively quickly (Hafner et al. 1994; Johnson et al. 2014) and resolving deeper level relationships within this group has proven difficult (Light et al. 2010). The complete genome of a representative of this group, the human head/body louse (*Pediculus humanus*), has been published (Kirkness et al. 2010) with an annotated protein-coding gene set which provides potential orthology targets for novel louse genome sequences. Finally, the genome size of sucking lice is relatively small (100-150 Mbp), making sequencing the entire genome through shotgun methods currently cost effective. We use this group as a test case for aTRAM with the goal of constructing a phylogenomic dataset of single copy orthologs from genome sequence data across divergent taxa.

METHODS

Phylogenomics of Sucking Lice (Anoplura) Using aTRAM

DNA was extracted using either a phenol chloroform extraction, QIAamp micro kit (Cat no. 56304; Qiagen, Hilden, Germany), or Zymo Genomic DNA-Tissue MicroPrep kit (Zymo Research, USA), and the lice were crushed to maximize DNA yields. (At the time of this study

library preparation kits typically required >200 ng DNA, such that some pooling of specimens was required to obtain sufficient DNA quantities. However, recent advances in gDNA library preparation allow quantities as low as 5 ng, such that single specimens can be used, unpub. data). The gDNA from each louse extract was sonicated using the Covaris M220 instrument to an average fragment size of 300-450bp (actual range was 200-600bp). The sheared gDNA was prepared for next-generation sequencing using TruSeq DNAseq or Kapa Library preparation kits. The resulting library was sequenced on either 1/3,1/2 or a full Illumina HiSeq2000 or 2500 using the TruSeq SBS sequencing kit v.1-2 for 101 or 161 cycles. All samples were sequenced paired-end, with 100 or 160bp reads. These genomic datasets are deposited in the NCBI:SRA (NCBI: SAMN03360966 – SAMN03360971; SAMN05930900 – SAMN0530910). The resulting reads were first examined using FastQC v0.10.1 (Babraham Bioinformatics) to screen for significant irregularities. The FASTX Toolkit v0.0.14 (Hannon Lab) was used for all quality trimming steps. For libraries sequenced with IlluminaHISeq version v1.8 or greater the first 3 bases were removed from the 5' end of the sequence read as they consistently had a lower quality score than the following nucleotides. All sequence reads were quality trimmed from the 3' end to remove bases with a phred score less than 28 using a sliding window of 1nt. Finally, any trimmed reads with fewer than 75nt in length were removed from the dataset (Supplementary Table 1).

The aTRAM software was designed to massively parallelize gene assembly from NGS data (Allen et al. 2015). aTRAM can run on a personal computer as well as a cluster with many processors. First, an aTRAM library is built where the paired-end reads are separated into smaller datasets called shards (Fig. 1a). Each shard contains the paired-ends for a subset of the entire dataset. For each shard, a BLAST formatted database is assembled from the first reads while their paired-ends are indexed. This process speeds up gene assembly in a few ways. First,

because there are known associations between paired-ends, read2 can be retrieved if read1 aligns to the target read, reducing the original BLAST search time. In other words, it is not necessary to spend processing time searching both pairs, rather, after if it is determined that read1 matches the locus, read2 can be incorporated into the assembly. Secondly, this process reduces the amount of time for each BLAST search by breaking the database into smaller BLAST libraries, as it is faster to search many smaller databases than a large one, even if the libraries are searched sequentially. Furthermore, splitting the entire short-read database into shards allows each to be searched in parallel, provided the user has access to multiple processors, and because each gene is assembled independently, many genes can be quickly assembled at one time. It is also possible to search only a fraction of the aTRAM library if desired. For example, if a high copy gene is the target (e.g. a mitochondrial gene) perhaps searching only half the library will produce a sufficient number of reads to assemble the gene. Therefore, the user can choose to search a fraction of the database and speed up the assembly process.

After an aTRAM library is built, the next step is to assemble the locus of interest. To do this, aTRAM uses an iterative process to BLAST a reference sequence (DNA; BLASTN or AA; TBLASTN) against the aTRAM library. The reference sequence can be from either a distantly or closely related taxon. The reads that match the reference and their mates are assembled using a *de novo* assembler (Fig. 1b). In the second iteration, the reference sequence is replaced by the assembled contigs and they are blasted against the aTRAM library (BLASTN; DNA to DNA). The previously assembled contigs along with any new matching reads are used in the *de novo* assembly to extend the contig. This process is repeated for a pre-defined number of iterations or until no new sequences are found. Because after the first iteration, the query sequences used in the BLAST are the assembled contigs from the same library, it is expected that more reads will

match and be assembled in the later iterations. This process helps to overcome any divergence issues with a distantly related initial reference sequence. After all the iterations are finished, the assembled contigs are scored by blasting back to the reference and a 'best' file is produced with the highest scoring contigs from each iteration. Therefore, each assembly will produce at least the same number of contigs as iterations. Because a longer sequence is used in the BLAST searches of subsequent iterations, the contig typically grows at each iteration and often the largest contig is assembled in the final iteration.

An aTRAM library was built from each of the 15 louse quality-trimmed short read datasets. A target gene phylogenomic dataset was identified using OrthoDB (Waterhouse et al. 2013). Orthologous 1:1 single copy genes were identified across the following 9 insect taxa: *Culex pipiens, Aedes aegypti, Anopheles gambiae, Drosophila melanogaster, Tribolium castaneum, Nasonia vitripennis, Apis mellifera, Pediculus humanus,* and *Acyrthosiphon pisum.* Ultimately 1,107 candidate genes were identified and first described in Johnson et al. (2013). aTRAM was then used to assemble these genes using the protein coding sequence from the reference louse genome *Pediculus humanus*. For each gene, aTRAM was set to run for three iterations and Trinity (Grabherr et al. 2011) was used as the *de novo* assembler. All of the contigs from the 'best' file were used in the post-processing steps described below.

For the purposes of this phylogenomic study, the intron sequences are expected to be highly variable and pose alignment problems across the deeply divergent taxa. Therefore, we developed a pipeline to extract the exon sequence from the aTRAM contigs (code available: https://goo.gl/mMzcEz). In the first step of this pipeline, the location of exon sequences are identified using the reference guided annotation software Exonerate v2.2.0 (Slater and Birney 2005). Those locations are then used to concatenate the identified exons into a single contig

containing only exon sequences (ie. codingDNA, herein cDNA), a process we call "stitching". In this process, exons are stitched together in two steps because, in cases where intron sequences are long, it is possible that aTRAM did not assemble completely through the intron sequences, particularly with only three iterations. Therefore, a few high scoring contigs may be assembled, each with different exons. Furthermore, because the beginning and end of exon sequences can be similar, Exonerate may annotate the same positions in the reference sequence on two separate contigs. To overcome these methodological issues, the pipeline we developed first identifies annotated contigs allowing for a small amount of overlap between them and stitches them together. These newly stitched contigs containing the exon sequences of interest are then re-annotated with Exonerate. In this second step, the exon sequences are found on a single contig therefore Exonerate is less likely to annotate the same amino acid position twice. For our 15 taxon dataset, we ran the 'exon-stitching' pipeline using the original reference amino acid sequences from *P. humanus* to annotate the exons.

As a further check of orthology, we subjected each final exon-containing contig to reciprocal-best-BLAST. For this dataset, we started with target sequences that were identified as 1:1 single copy orthologs across a wide diversity of insects, so we expected that these sequences would have a high probability of being orthologs for all the taxa in our study. However, to verify, we blasted the resulting cDNA sequences against the *P. humanus* proteome to determine if the original target ortholog was returned as the top hit as in Allen et al. (2015). Because the original *P. humanus* sequence was used as the target for BLAST in the aTRAM gene assembly, blasting the cDNA contigs back against the *P. humanus* proteome provides the reciprocal BLAST for the test. Any cDNA contigs that did not pass this reciprocal-best-blast test were removed from the dataset.

The cDNA sequences recovered from the above procedure were categorized as either full-length (assemblies representing 50% or more of the cDNA length of the reference gene) or fragmentary (assemblies representing less than 50% of the cDNA length of the reference gene). The cDNA sequences were translated into amino acid sequences using the Exonerate annotation data to identify the appropriate frame. The full-length amino acid sequences were aligned using PASTA v1.6.3 (Mirarab et al. 2015). The remaining fragmentary sequences were then inserted into the PASTA alignment using UPP v2.0 (Nguyen et al. 2015), a method designed for accurately aligning fragmentary data into an existing reference alignment.

A partitioned maximum likelihood (ML) analysis was then performed to better handle rate heterogeneity across genes. The protein alignments were back-translated into cDNA alignments, and three alignment files were created, one for each codon position. The distribution of the GC content for each nucleotide position was plotted to check for nucleotide composition bias for each position. The GC content for the third codon position showed strong bias, and thus, was omitted from subsequent phylogenetic analyses. Next, RAxMLv8.1.3 (Stamatakis 2014) was run under GTR+GAMMA to estimate the GTR rate matrix and alpha parameter for the first and second codon positions of each gene. The rate matrices were combined into a single matrix, where each row was a codon position for a particular gene, and each column was either one of the entries in the GTR rate matrix or the alpha shape parameter. The columns in the matrix were centered and a PCA was conducted using R (R Core Team, 2013). In order to group codon positions with similar rate matrices into the same partition, k-means clustering was performed on the points on the PCA plot using R. We varied the number of partitions between 2 and 30 to examine the impact on the ratio of the between cluster sum of squares and the total sum of squares.

Phylogenetic trees were built using both a concatenated dataset as well as gene tree-species tree coalescence analysis (Supplemental Datasets). First, the k-means clustering results were used to partition the concatenated alignment. A maximum likelihood (ML) species tree was estimated using a partitioned RAxML analysis under GTR+GAMMA. Branch support was estimated using 100 bootstrap replicates. Second, the gene trees were estimated using RAxML under GTR+GAMMA on the individual gene alignments with the third codon position removed. Branch support was estimated on each of the gene trees using 100 bootstrap replicates. ASTRAL II v4.9.9 (Mirarab et al. 2015), a coalescence-based analysis, was performed to estimate a species tree from the gene trees. Local branch support on the ASTRAL tree was estimated using quartet frequencies of the gene trees (Sayyari and Mirarab 2016).

Testing aTRAM with varying sequence coverage and genetic divergence

To determine both coverage and diversity thresholds for successfully applying aTRAM to a genomic dataset, we selected seven taxa representing closely and distantly related species to the reference sequence. Three of these taxa were included in the original louse dataset, and we added four additional taxa to represent deeper divergences (0 – 145 million years; Table 1). The coverage of each dataset was calculated using three genes selected at random and assembled with aTRAM. These genes were then BLASTed against the entire aTRAM library and the hits recovered and aligned to the aTRAM-assembled contig. Coverage was then calculated by averaging the number of reads at five points across the gene. Using the fraction feature in aTRAM, each database was then manipulated to represent five test coverages (1X, 5X, 10X, 20X and 40X). For each level of coverage, 100 genes were assembled with aTRAM and the number of successful assemblies counted (i.e. genes producing a “best” file).

RESULTS

Phylogenomics of Sucking Lice (Anoplura) Using aTRAM

Genomic libraries for Anoplura species and outgroups produced between 42 and 154 million reads for each taxon, resulting in between 4,295 and 27,907 Mbp of DNA sequence per taxon. Given the reported genome size of *P. humanus* is 108 Mbp, (Kirkness et al., 2010) and assuming similar genome sizes for other species of Anoplura, we would expect coverage to be between 30–300X. This level of coverage is similar to what we found by mapping reads against test target contigs (see below), so it appears that genome size across Anoplura is relatively stable (~100–150 Mbp).

Of the 1,107 potentially orthologous genes, 98% of them produced aTRAM assemblies for all taxa (Table 2). Exonerate annotation and subsequent exon concatenation resulted in variable coverage of the protein coding portions for these genes as compared to the reference amino acid sequence. For all taxa, 78% of the genes were over 50% complete with respect to the target sequence. Between 14 and 50 (mean 27) genes, depending on taxon, did not pass the reciprocal best BLAST test in terms of returning the original target gene. Following these steps, six genes either had fewer than 4 sequences or contained all identical sequences, and these were removed, leaving 1,101 genes in the final dataset for the alignment (Table 2). The reference amino acid lengths of these genes varied from 53 to 4,210 with a mean of 548. The sum of the lengths for the target, *P. humanus*, for these genes was 1.78 million bp. Of the 1,107 genes 56% contained at least one fragmentary sequence (i.e. < 50% complete). The number of fragments per gene ranged from 6 to 20% (mean 12%). After alignment with PASTA and UPP, the total number of aligned base pairs was 3.08 million.

The base composition of the third codon position was highly variable across taxa, ranging between 25% and 50% GC. In contrast, the base composition of the first and second positions across taxa was much more stable (Fig. 2). To avoid base composition biases affecting phylogenetic analysis, we removed the third codon position from further analyses. When we computed the rate parameters for the codon positions, we found that 182 codon positions had a rate parameter that was more than 10 standard deviations from the mean value; these were excluded from the data matrix (Supplemental Figure1). For partitioning the data matrix for likelihood analysis, the 90% threshold for the k-means clustering suggested 12 partitions. Thus, first and second codon positions were grouped together into 12 partitions and likelihood model parameters estimated separately for each partition, with codon positions with extreme rate parameters excluded. In total 2,020 codon positions from 1,080 genes were used to generate the final concatenated partitioned data matrix.

The RAxML analysis of the concatenated, partitioned data matrix resulted in a single, completely resolved tree, with all nodes supported in 100% of bootstrap replicates (Fig. 3). The species tree obtained from the ASTRAL gene tree analysis was identical to the tree from the concatenated matrix and all nodes were supported with 100% local support.

In this tree, the lice from primates formed a monophyletic group (*Pedicinus, Pthirus, Pediculus*), as did the lice from pinnipeds (*Antarctophthirus* and *Proechinophthirus*), suggesting concordance with host relationships for those groups (Fig. 3). However, incongruence was found with the ungulate louse (*Linognathus*), it was embedded within the lice of rodents (*Hoplopleura* and *Neohaematopinus*). The only previous molecular phylogenetic study of the relationships among the major lineages of sucking lice (Light et al. 2010) is consistent with our results. However, this prior study was based on only three genes from Sanger PCR targeted sequencing

and did not confidently place the pig louse *Haematopinus*. Here we find *Haematopinus* to be the sister to the pinniped lice with 100% support. Thus, our tree based on 1,102 genes assembled from across the genome was able to increase resolution within this group and also greatly increase the support for previous results.

Testing aTRAM with Varying Sequence Coverage and Genetic Divergence

We evaluated the impact of genome coverage and genetic distance on the success rate and completeness of aTRAM assemblies. To do this, we assembled 100 genes experimentally varying the coverage from 1X to 40X across seven taxa varying in divergence times from the target, *P. humanus*, (0-145 my), and including *P. humanus*. Additional taxa, from the same genus, same suborder, different suborders, and a free-living bark louse were also analyzed, with an estimated maximum divergence time of 145 million years ago (*Stimulopalpus*; Smith et al. 2011; Table 2). At levels of coverage 10X and higher, few assembled genes are lost with increasing distance from *P. humanus* (Fig. 4). Even outside of parasitic lice (Phthiraptera), only a few of the 100 test genes were not assembled. This result suggests that even highly divergent reference sequences (at least up to 150 Ma.) can be used as targets for fairly low coverage aTRAM assemblies. Reducing genome coverage to 5X or lower resulted in the majority of the genes not assembling in the distantly related taxon. However, for the closely related taxa (0-5my) 74 and 66, respectively, of 100 genes assembled at 5X coverage (Fig. 4). These results indicate that aTRAM is robust to both highly divergent reference sequences with moderately low coverage (5-10X) and to very low coverage (1-5X) for closely related taxa.

DISCUSSION

We generated whole genome datasets from a single Illumina short read library for each of 15 taxa (100-160bp paired-end reads from 500bp insert libraries), with modest genome coverage

(as low as 30X), making this approach economically feasible for a variety of taxa with modest (< 1000 Mbp) genome sizes. Over 1,000 genes were assembled from each taxon representing a dataset with 97.2% gene completeness. After our post-processing exon stitching and reciprocal best BLAST filtering, only six genes were removed, producing a final dataset that was 99% complete. The resulting concatenated data matrix totaled ~1.29 million base pairs (after removing the 1st and 2nd codon positions with extreme rate parameters, Supplemental Dataset). The maximum-likelihood phylogeny derived from this dataset was completely resolved and all branches were supported with 100% of bootstrap replicates. In addition, ASTRAL gene tree analyses resulted in an identical tree, with all branches supported at 100%. The tree estimated from this phylogenomic dataset was largely congruent with a previous study (Light et al. 2010) of sucking lice (Anoplura) based on targeted Sanger sequencing of three genes. However, we were able to confidently resolve the position of a difficult taxon, *Haematopinus*, while Light et al. (2010) recovered it as part of a polytomy at the base of the tree. With respect to the taxa sampled in the current study, the phylogenetic results are not congruent with the morphological analysis of Kim (1988), which is not surprising, as lice are known to undergo convergent morphological evolution. In addition to increased phylogenetic resolution, we were able to substantially increase branch support for the tree over the Light et al. (2010) study.

aTRAM Guidelines and Findings

In general, aTRAM works by querying a short read genomic database for reads matching a reference sequence. These reads are assembled locally, removing much of the difficulty of full genome assembly. Whole genome assembly normally considers all the reads simultaneously and can add errors in the process, making the assembly more computationally complex. Here we used TBLASTN, which uses an amino acid query. The advantage of this approach is that

matches can be found in taxa that are highly divergent from the target (>100 Myr), unlike reference based assembly (read-mapping) approaches, which require more similar reference sequences. In our evaluation of the sensitivity of aTRAM to increasing divergence from the target, we found that there was only a slight drop in performance with divergences up 150 mya, provided the coverage was 10X or greater. We expect that TBLASTN searches will reach a limit with even deeper divergences, because it is more difficult to find significant hits among short (<160 bp) reads. In addition, the success of TBLASTN may also be limited by exon size, with very short exons less likely to produce significant hits, given that a greater fraction of the reads covering the exon will have part of their sequence contained within the intron. It is also possible to assemble the entire open reading frame (introns and exons) or other gene regions (e.g. UCEs from targeted sequencing) using aTRAM (unpub data), but in this case BLASTN is needed. The sensitivity of BLASTN searches in aTRAM to increasingly divergent references has not been evaluated; however, it is likely that sequence divergence will be more of an issue with DNA-DNA BLAST searches. Still there may be ways to optimize the aTRAM run for these datasets, including setting different e-values for the BLASTN searches.

We found that aTRAM is sensitive to decreased coverage. The assembly success rate decreased at around 10X coverage, however 80 – 90% of the genes were still assembled. Below 5X coverage, the assembly rate dropped rapidly for all but the two most closely related taxa. In part, this may be due to the variance in coverage, such that some gene regions have no coverage at these levels. However, some of this drop may be due to limitations of the assemblers that are incorporated into aTRAM (ABySS, Simpson et al. 2009; Trinity, Grabherr et al. 2011; and Velvet, Zerbino 2011). In the future, incorporating additional assembly software may provide successful assemblies at even lower coverages. Some preliminary analysis suggests Velvet is

more likely to assemble loci with very low coverage (1-2X), while ABySS requires higher coverage levels. However we did not compare the error rates between the assemblers. To ensure the best results, at least 10X sequencing coverage seems to be a good goal when using aTRAM, particularly with more divergent taxa.

For this dataset we ran aTRAM using only three iterations because we were ultimately focused on exon sequences for phylogenetic analysis. Because the original target sequence included all the exon regions, we expected that the majority of those would assemble in a few iterations. While we did not examine the success rate of intron assembly at three iterations, many full genes were found on a single contig, suggesting the intron sequences were intact. However, for datasets where introns are of interest, it may be necessary to run more iterations to ensure that the introns are completely assembled. For this dataset, we used all the contigs printed to the 'best' file (those with the best BLAST score to the reference sequence) for downstream analysis. because we expected that some genes may assemble with different exons on different contigs. If we had used more iterations, all of the contigs may have been joined together after the introns had been completely assembled.

Comparison with Other Approaches

In this study, we make use of complete genome sequencing to assemble genes relevant for phylogenomic analysis. In the case of taxa with very large (>2000 Mbp) genomes, it may not yet be economically feasible to sequence the genomes of a large number of taxa using this approach. Two popular genome reduction approaches, transcriptome sequencing (e.g., Misof et al. 2014) and target capture (e.g., UCEs [Faircloth et al., 2012, 2015] and AHE [Lemmon et al. 2012]) are now being widely applied in phylogenomics. These approaches have a cost savings

advantage over complete genome sequencing for taxa with large genomes; however, they also have limitations.

For transcriptome sequencing, a large quantity of high quality material for RNA extraction is needed. For many large-scale phylogenetic studies, particularly for small-bodied organisms, this is not possible. Many taxa are rare or difficult to obtain and preservation of existing material is not compatible with successful RNA extraction. In addition, transcriptome sequencing is limited to expressed genes, resulting in increased potential for missing data (depending on the ortholog set) because some genes are not expressed at all life stages. Also, sequencing is limited to exon portions of the genes, and for some phylogenetic problems, such as among closely related species, intron sequence may be more informative (Johnson and Clayton 2000; Hackett et al, 2008; Jarvis et al., 2014; Mirarab et al., 2014).

Target capture approaches have the advantage over transcriptome sequencing in that DNA is directly sequenced rather than RNA. Since DNA is more stable than RNA, a wider variety of preserved materials can be used. These approaches work by binding genomic DNA to a variety of oligoprobes and preferentially sequencing the DNA bound to these probes. This approach reduces the fraction of the genome that is sequenced and further reduces sequencing costs by multiplexing. However, one disadvantage of these approaches is that a probe set needs to be designed in advance of the study, requiring prior genomic knowledge of the taxa or a close relative. In addition, there is often mixed success of probe binding rates across taxa in the study (Hedke et al. 2013), resulting in data matrices with a relatively large fraction of missing data, particularly if the clade of interest has variable rates of evolution among the members. Because only a fraction of genomic DNA is sequenced, these approaches also typically require more

starting DNA than sequencing the entire genome directly (Ekblom and Wolf 2014; Faircloth et al. 2015), which potentially limits the application for very small-bodied organisms.

Fundamentally, the main disadvantage of genome reduction methods is that further analyses are limited by the data collected for the initial study. Future phylogenetic studies adding taxa will have to collect data for the same loci. While transcriptomes may be more flexible in this regard, because ortholog sets can be generated after sequencing, this can be a major limitation for target capture approaches.

Whole genome sequencing approaches, followed by aTRAM gene assemblies, are much more flexible. Additional genes can be later added to analyses and sequencing efforts in principle only need to be conducted once. Existing data can later be mined for genes of interest. In addition, data from other associated genomes (e.g. mitochondria and symbionts) are also often obtained at the same as the genome of interest. Data from these genomes can be used in additional studies, which maximizes the use of the data collected. For example, the genomic data used in this study has also been used to assemble complete symbiont genomes (Boyd et al. 2014, 2016).

Limitations of Whole Genome Sequencing and aTRAM

While complete genome sequencing followed by aTRAM gene assembly can be a highly attractive option for generating phylogenomic datasets, there are some limitations that must be considered. First, sequencing the whole genome is not yet particularly cost effective for organisms with very large (>1000 Mbp) genomes. Related to cost efficiency, our tests revealed a certain amount of locus coverage (around 10X) is needed to have a high success rate of gene assembly, particularly for more divergent taxa. However, further optimization of assembly algorithms may eventually allow the use of lower coverage. Additionally, nuclear genome

coverage may be swamped by high-copy organellar genomes, further reducing the effective coverage of nuclear loci (Kane et al 2012).

Another important consideration is the availability of reference target sequences. While complete genome assemblies and annotations are becoming more readily available for a wide variety of organisms, some groups may not yet have a suitable reference for identifying target loci. In addition, the post aTRAM annotation of exons using Exonerate depends on having a well-annotated reference, because the Exonerate annotation relies on accurate amino acid sequences. Genome annotation is still a major enterprise for most genome studies and even well studied genomes are being re-annotated as methods and information improves (e.g. Camus et al., 2002; Darwish et al, 2015). For studies that use entire gene regions, not just protein coding portions, accurate annotation of the target gene is not as crucial, because in aTRAM the assembly itself is not done with respect to the target, only retrieval of the reads to assemble.

Future Directions

The aTRAM approach has the potential for considerable extension and expansion in the future. While the current study focused on orthologs, aTRAM also has the possibility to assemble paralogs within gene families. For protein coding genes that are part of a gene family, TBLASTN searches are likely to recover reads from all the members of the gene family. We expect that the assembly algorithms will assemble these paralogous genes into separate contigs. These gene paralogs can be analyzed phylogenetically in a gene tree approach to identify gene duplication and loss events. Preliminary investigation of this approach indicates it should be possible (unpub data) to separately assemble gene duplicates within a species using aTRAM.

aTRAM will likely prove useful for other types of assemblies and datasets. For example, preliminary data suggests aTRAM can be used to successfully assemble UCEs from targeted

sequencing datasets, and because aTRAM contigs grow at each iteration, it may be ideal for assembling into the variable regions surrounding UCEs. Furthermore, small circular genomes may be targeted including chloroplast and mitochondrial datasets. However, further testing is needed and different parameters may be optimal for these types of assemblies.

Another line of investigation to pursue is using aTRAM to develop a reference gene set and then aligning reads from other related taxa with reference-based assembly, such as Bowtie2, BWA, or SOAPaligner (Li and Durban 2009; Li, et al., 2009; Lagmead and Salzberg, 2012). This may save on cost and still allow full genome DNA sequencing. For example, only a single species would be sequenced at higher coverage (>10X) for aTRAM contig assembly and the remaining closely related species could be sequenced at a lower coverage for read-mapping. Because read alignment can produce consensus sequences at low coverage, this approach may be more economically feasible than the relatively higher coverage for genome sequences needed for aTRAM.

SUPPLEMENTAL INFORMATION

All data files and supplementary documents can be found in the Dryad data repository http://datadryad.org, http://dx.doi.org/10.5061/dryad.26j38

Supplementary Table: DNA extraction, and quality clean up for each dataset. Illumina reads. Alignments of each gene and the tree analysis.

Supplementary Figure: Box plot of the standard deviations away from mean for each codon position for each of the GTR rate parameters. The majorities of the extreme outliers fell above 10 standard deviations from the mean and were removed from the analysis.

Supplementary Dataset 1: Alignment and phylogenetic tree of the concatenated 1,101 gene alignment.

Supplementary Dataset 2: All 1,101 gene trees and alignments for the 15 taxon dataset.

ACKNOWLEDGEMENTS

This research was supported in part by NSF grants DEB-0612938, DEB-1050706, DEB-1239788, and DEB-1342604 to KPJ, NSF XSEDE DEB-160002 to KPJ, BMB, and JMA, DBI-1461364 and ABI-1458652 to TW, and NSF DEB1310824 to BMB. 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. *Hoplopleura arboricola* (MSB Para 20350) and *Neohaematopinus pacificus* (DMNS ZM.11420) specimens were acquired from the Denver Museum of Nature & Science and the Museum of Southwestern Biology. Samples of *Antarctophthirus microchir* (CNP-PAR 82) were collected by M.S. Leonardi with funds from Marine Mammals Lab (CENPAT).

REFERENCES

Alkan C., Sajjadian S., and Eichler E.E. 2011. Limitations of next-generation genome sequence assembly. Nat. Methods 8:61-65.

Allen J.M., Huang D.I., Cronk Q.C., Johnson K.P. 2015. aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. BMC Bioinformatics 16:98.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Boyd B.M., Allen J.M., de Crécy-Lagard V., Reed D.L. 2014. Genome sequence of Candidatus Riesia pediculischaeffi, endosymbiont of chimpanzee lice, and genomic comparison of recently acquired endosymbionts from human and chimpanzee lice. G3: Genes Genom. Genet. 4:2189-2195.

Boyd B.M., Allen J.M., Koga R., Fukatsu T., Sweet A.D., Johnson K.P., Reed D.L. 2016. Two bacteria genera, Sodalis and Rickettsia, associated with the seal louse Proechinophthirus fluctus (Phthiraptera: Anoplura). Appl. Environ. Microbiol. 00282:-16

Camus J.C., Pryor M.J., Médigue C., Cole S.T. 2002. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiol. 148:2967-73.

Darwish O., Shahan R., Liu Z., Slovin J.P., Alkharouf N.W. 2015. Re-annotation of the woodland strawberry (Fragaria vesca) genome BMC Genom. 16:29 DOI: 10.1186/s12864-015-1221-1.

Ekblom R., Wolf B.W. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7:1026-1042.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Sys. Biol. 61:717–726.

Faircloth B.C., Branstetter M.G., Whites N.D., Brady S.G. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. 15:489-501.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–52.

Hackett S.J., Kimball R.T., Reddy S., Bowie R.C.K., Braun E.L., Braun M.J., Chojnowski J.L., Cox W.A., Han K., Harshman J., Huddleston C.J., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Steadman D.W., Witt C.C., Yuri T. 2008. A phylogenomic study of birds reveals their evolutionary history. Science (New York, N.Y.), 320(5884), 1763–8.

Hafner M., Sudman P., Villablanca F. 1994. Disparate rates of molecular evolution in cospeciating hosts and parasites. Science 265:1087-1090.

Hedtke S.M., Morgan M.J., Cannatella D.C., Hillis D.M. 2013. Targeted Enrichment: Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time. PLoS One. 8(7): e67908. doi: 10.1371/journal.pone.0067908

Howison M., Zapata F., Dunn C.W. 2013. Toward a statistically explicit understanding of de novo sequence assembly. Bioinformatics 29:2959–2963.

Howison M., Zapata F., Edwards E.J., Dunn C.W. 2014. Bayesian genome assembly and assessment by Markov Chain Monte Carlo sampling. PLoS ONE 9:e99497.

Kane N.C., Sveinsson S., Dempewolf H., Yang J.Y., Zhang D., Engels J.M.M., Cronk Q.C. 2012. Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am. J. Bot. 99: 320–329.

Kim K. C. 1988. Evolutionary parallelism in Anoplura and eutherian mammals. In: Biosystematics of Haematophagous Insects. Syst. Assoc. (V:37)

Johnson K.P. and Clayton D.H. 2000. Nuclear and mitochondrial genes contain similar phylogenetic signal for pidgeons and doves (Aves: Columbiformes). Molec Phylogen and Evol 14:141-151

Johnson K.P., Walden K.K.O., Robertson H.M. 2013. Next-generation phylogenomics using a Target Restricted Assembly Method. Mol. Phylogenet. Evol. 66:417–22.

Johnson K.P., Allen J.M., Olds B.P., Mugisha L., Reed D.L., Paige K.N., Pittendrigh B.R. 2014. Rates of genomic divergence in humans, chimpanzees and their lice. P. R. Soc. B. 281:20132174.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B.,

Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yingi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa F., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K-P., O'Brien S., Haussler, D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Kirkness E.F., Haas B.J., Sun W., Braig H.R., Perotti M.A., Clark J.M., Lee S.H., Robertson H.M., Kennedy R.C., Elhaik E., Gerlach D., Kriventseva E.V., Elsik C.G., Graur D., Hill C.A., Veenstra J.A., Walenz B., Tubío J.M.C., Ribeiro J.M.C., Rozas J., Johnston J.S., Reese J.T., Popadic A., Tojo M., Raoult D., Reed D.L., Tomoyasu Y., Krause E., Mittapalli O., Margam V.M., Li H-M., Meyer J.M., Johnson R.M., Romero-Severson J., VanZee J.P., Alvarez-Ponce D., Vieira F.G., Aguadé M., Guirao-Rico S., Anzola J.M., Yoon K.S., Stycharz J.P., Unger M.F., Christley S., Lobo N.F., Seufferheld M.J., Wang N., Dasch G.A., Struchiner C.J., Madey G., Hannick L.I., Bidwell S., Joardar V., Caler E., Shao R., Barker S.C., Cameron S., Bruggner R.V., Regier A., Johnson J., Viswanathan L., Utterback T.R., Sutton G.G., Lawson D., Waterhouse R.M., Venter J.C., Strausberg R.L., Berenbaum M.R., Collins F.H., Zdobnov E.M., Pittendrigh B.R. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. P. Natl. Acad. Sci. USA 107: 12168–73.

Langmead B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4): 357–359.

Lasken R.S., and Stockwell T.B. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. 7:19.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–44.

Li H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform Bioinformatics 25(14):1754-1760.

Li R., Yu C., Li Y., Lam T., Yiu S., Kristiansen K., Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (15): 1966-1967. doi: 10.1093/bioinformatics/btp336.

Light J., Smith V., Allen J. Durden L.A., Reed D.L. 2010. Evolutionary history of mammalian sucking lice (Phthiraptera: Anoplura). BMC Evol. Biol. 10:1.

Mirarab S., Reaz R., Bayzid M.S., Zimmerman T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree. Bioinformatics 30:i541–i548.

Mirarab S., and Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Proceedings ISMB 2015, and Bioinformatics 2015 31 (12): i44-i52 doi: 10.1093/bioinformatics/btv234

Mirarab S., Nguyen N., Guo S., Wang L.-S., Kim J., Warnow, T. 2015. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acidsequences. J. Comput. Biol. 22:377–86.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware, J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler t., Rust

J., Aberer A.J., Aspöck H., Bartel D., Blanke A., Berger S., Böhm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui Makiki F., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schütte K., Sekiya K., Shimizu Sh., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan Meihua T., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xun X., Zhang Y., Yang H., Want J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.

Morozova O., Hirst M., Marra M.A. 2009. Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genom. Hum. G. 10:135–51.

Niehuis O., Hartig G., Grath S., Pohl H., Lehmann J., Tafer H., Donath A., Krauss V., Eisenhardt C., Hertel J., Petersen M., Mayer C., Meusemann K., Peters R.S., Stadler P.F., Beutel R.G., Bornberg-Bauer E., McKenna D.D., Misof B. 2012. Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera. Curr Biol 22(14):1309-1313.

Nguyen N.D., Mirarab S., Kumar K., Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. Genome Biol. 16:1-15.

Pop M. 2009. Genome assembly reborn: recent computational challenges. Brief. Bioinform. 10:354-366

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Salzberg S.L., Yorke J.A. 2005. Beware of mis-assembled genomes. Bioinformatics 21: 4320–4321.

Sayyari E., and Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Molec. Biol. Evol. 10.1093/molbev/msw079.

Simpson, J.T., Wong K., Jackman S.D., Schein J.E., Jones, S.J.M., Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research, 19(6), 1117–23.

Slater G.S.C., Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Smith V.S., Ford T., Johnson K.P., Johnson P.C., Yoshizawa K., Light J.E. 2011. Multiple lineages of lice pass through the K-Pg boundary. Biol. Letters 7:782-785.

Snodgrass R. 1944. The feeding apparatus of biting and sucking insects affecting man and animals. Smithsonian Miscellaneous Collections 104:1-113.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England), 30(9), 1312–3. http://doi.org/10.1093/bioinformatics/btu033

Waterhouse R.M., Tegenfeldt F., Li J., Zdobnov E.M., Kriventseva E.V. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41:D358–365.

Wetterstrand K.A. 2013. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). National Human Genome Research Institute

Zheng Z., Liebers M., Zhelyazkova B., Cao Y., Panditi D., Lynch K.D., Chen J., Robinson H.E., Shim H.S., Chmielecki J., Pao W., Engelman J.A., Iafrate A.J., Le L.P. 2014. Anchored multiplex PCR for targeted next-generation sequencing. Nat. Med. 20:1479–84.

Zerbino D.R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Current Protocols in Bioinformatics 11:11.5.

FIGURE CAPTIONS

FIGURE 1: Details of the aTRAM process. **a.** Creating an aTRAM library from paired-end sequences from NGS sequencing. First, DNA reads are sequenced with paired-end technology, split into shards and all read1 sequences are formatted into a blast database and the corresponding paired-end (read2) indexed. **b.** assembling loci with aTRAM. First a reference sequence is used to BLAST against the aTRAM library, the matches and their paired-ends are found and assembled with a *de novo* assembler. The assembled contigs are then used as the reference sequence in the next iteration, and the process repeats. Figure modified from Allen et al., 2015.

FIGURE 2: Base composition (GC%) by codon position (first, second, third) across 15 target taxa. The third codon position was more variable across the taxa and was removed for the tree building steps.

FIGURE 3: Concatenated RAxML tree of 1,101 genes of bird and mammal lice (Phthiraptera). Hosts are represented on the nodes by the silhouettes. Values on the nodes represent bootstrap values. All nodes were supported with 100% bootstrap in the concatenated as well as gene-tree species tree analysis with ASTRAL.

FIGURE 4: Coverage and diversity results from aTRAM. Taxa with warmer colors (more red) are more closely related to the reference taxon, *Pediculus humanus*, while taxa with cooler colors (more blue) are more distantly related. Overall, both coverage and divergence from the reference affected the number of genes assembled. At higher coverages all genes were successfully assembled across all taxa. However, more genes were assembled from taxa that were more closely related to the reference at lower coverage levels. Between 10 to 5X coverage distance

from the reference affected the success of gene assembly and lower than 5X coverage the divergent taxa had a major drop in assembly rate. This suggests that 10X coverage is a good target for more divergent assemblies, while more closely related assemblies 5X coverage might still assemble the majority of the genes.

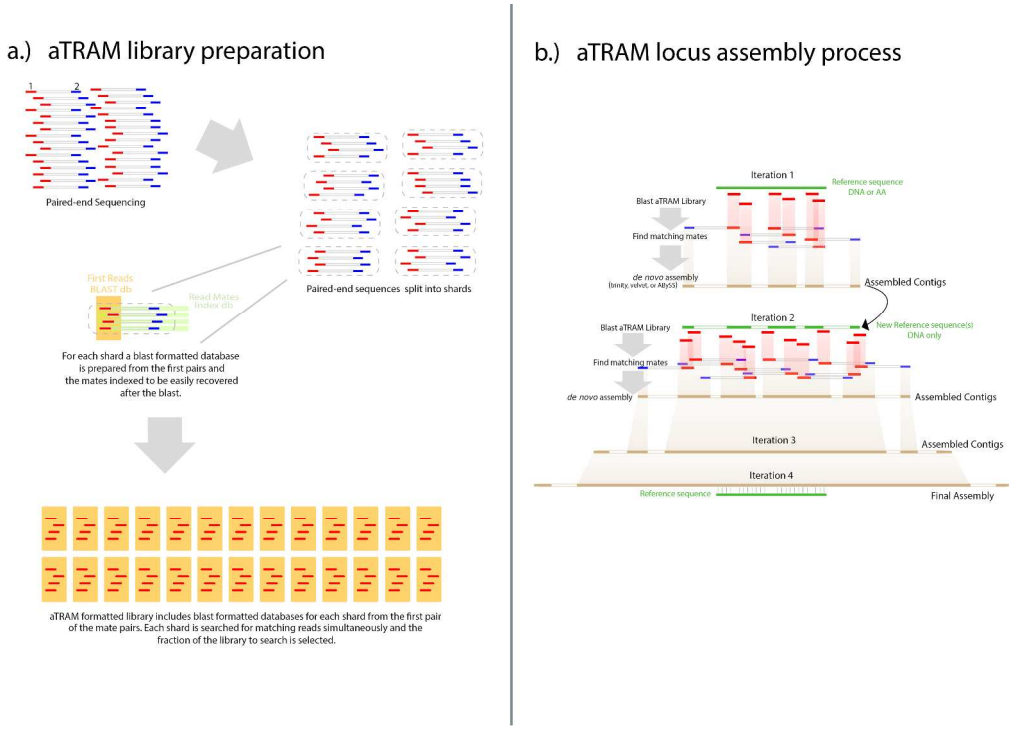


FIGURE 1: Details of the aTRAM process. a. Creating an aTRAM library from paired-end sequences from NGS sequencing. First, DNA reads are sequenced with paired-end technology, split into shards and all read1 sequences are formatted into a blast database and the corresponding paired-end (read2) indexed. b. assembling loci with aTRAM. First a reference sequence is used to BLAST against the aTRAM library, the matches and their paired-ends are found and assembled with a de novo assembler. The assembled contigs are then used as the reference sequence in the next iteration, and the process repeats. Figure modified from Allen et al., 2015.

415x299mm (300 x 300 DPI)

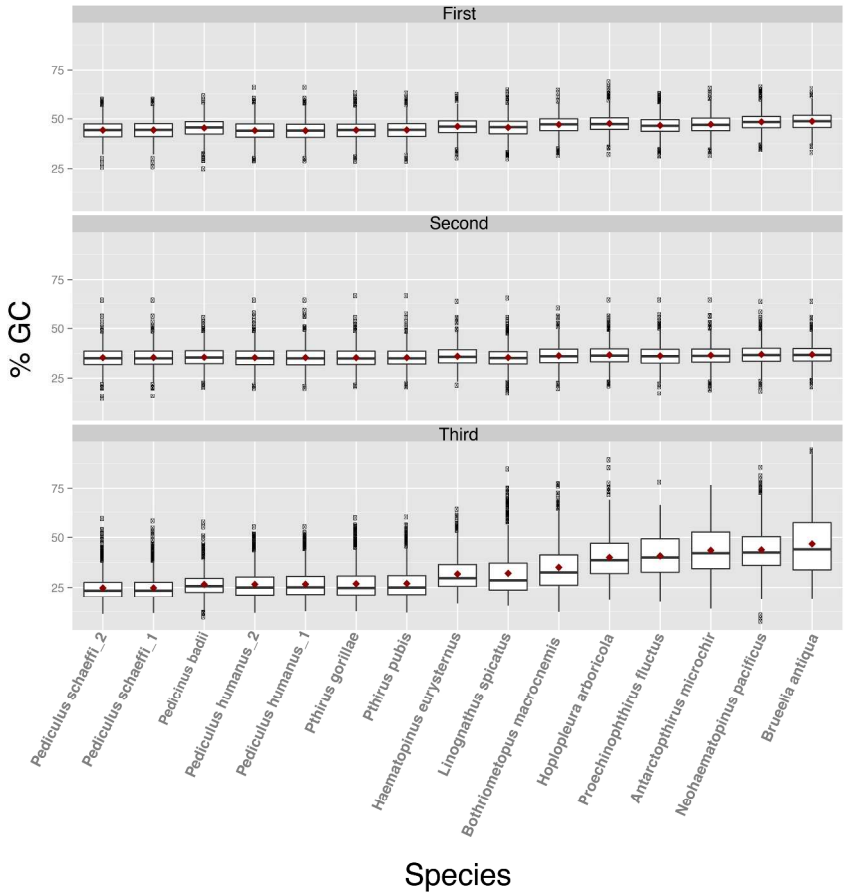


FIGURE 2: Base composition (GC%) by codon position (first, second, third) across 15 target taxa. The third codon position was more variable across the taxa and was removed for the tree building steps.

304x277mm (300 x 300 DPI)

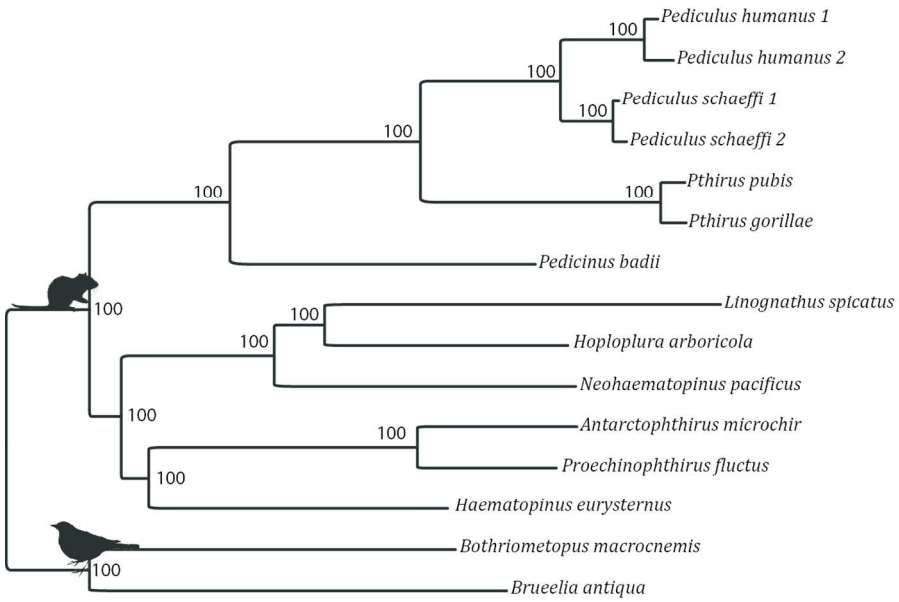
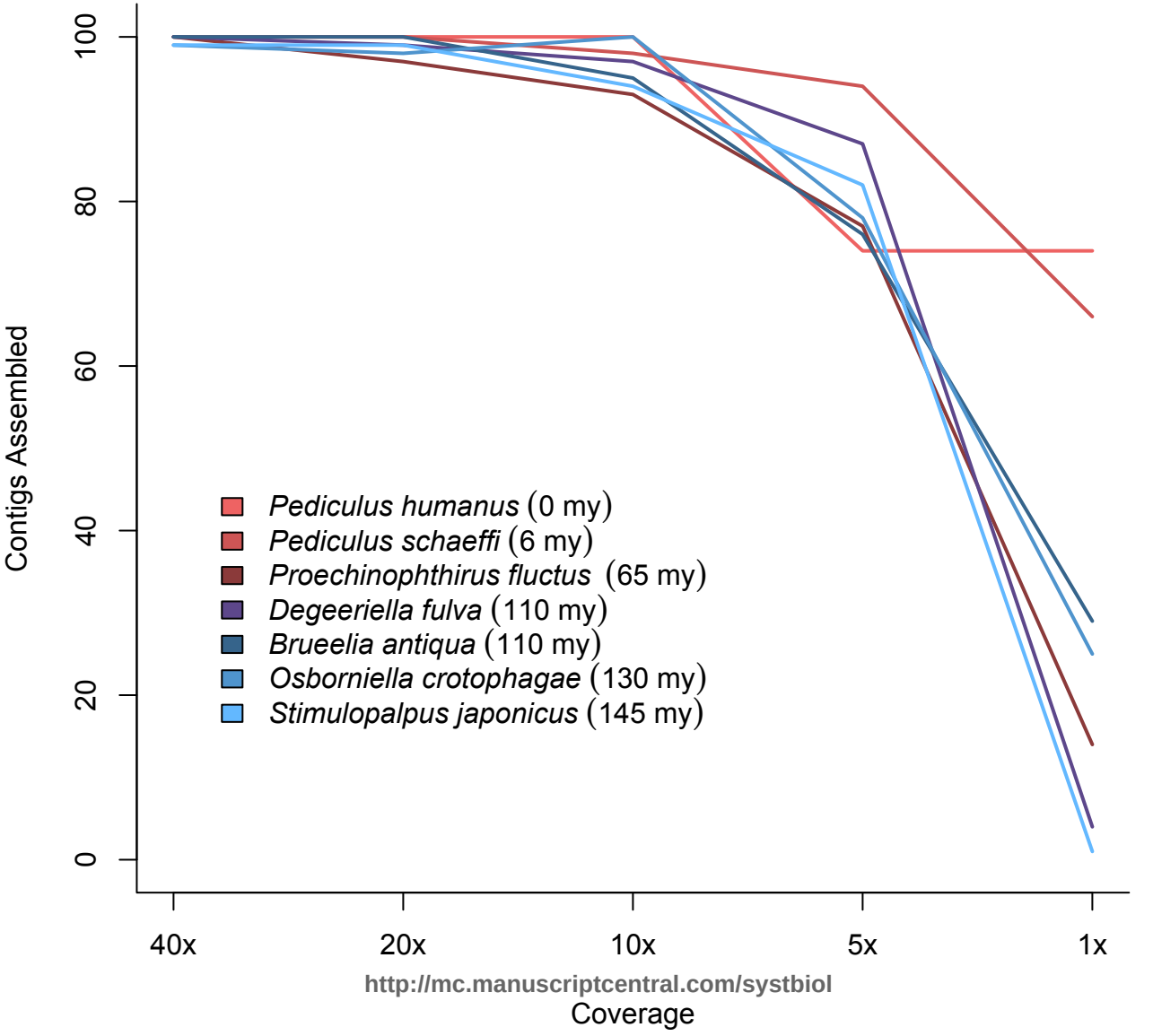


FIGURE 3: Concatenated RAxML tree of 1,101 genes of bird and mammal lice (Phthiraptera). Hosts are represented on the nodes by the silhouettes. Values on the nodes represent bootstrap values. All nodes were supported with 100% bootstrap in the concatenated as well as gene-tree species tree analysis with ASTRAL.

255x154mm (150 x 150 DPI)


Contigs Assembled
0
20
40
60
80
100
40x
20x
10x
5x
1x
Coverage
Pediculus humanus (0 my)
Pediculus schaeffi (6 my)
Proechinophthirus fluctus (65 my)
Degeeriella fulva (110 my)
Brueelia antiqua (110 my)
Osborniella crotophagae (130 my)
Stimulopalpus japonicus (145 my)

TABLES

TABLE 1: Seven Taxa Selected for Divergence and Coverage Test

	Louse Species	Family	Suborder	Order	SRA	Host	Estimated Divergence Time
Reference	***Pediculus humanus***	**Pediculidae**	**Anoplura**	**Phthiraptera**	AAZO00000000.1	***Homo sapiens***	**--**
1	*Pediculus humanus*	Pediculidae	Anoplura	Phthiraptera	SAMN05930905	*Homo sapiens*	0
2	*Pediculus schaeffi*	Pediculidae	Anoplura	Phthiraptera	SAMN02438447	*Pan troglodytes*	6 Ma.
3	*Proechinophthirus fluctus*	Echinophthiriidae	Anoplura	Phthiraptera	SAMN03360968	*Callorhinus ursinus*	65 Ma.
4	*Degeeriella rufa*	Philopteridae	Ischnocera	Phthiraptera	SAMN05930901	*Falco berigora*	110 Ma.
5	*Brueelia antiqua*	Philopteridae	Ischnocera	Phthiraptera	SAMN03360970	*Catharus ustulatus*	110 Ma.
6	*Osborniella crotophagae*	Menoponidae	Amblycera	Phthiraptera	SAMN05930903	*Crotophaga ani*	130 Ma.
7	*Stimulopalpus japonicus*	Amphientomidae	Troctomorpha	Psocodea	SAMN05930910	*Free living*	145 Ma.

TABLE 1: Taxa represented in the divergence and coverage test. For each taxon 100 genes were assembled with aTRAM from each of the following estimated coverages 40X, 20X, 10X, 5X and 1X using the fraction feature in aTRAM to manipulate the coverage level from each dataset. These taxa were selected for their representative divergences in time (Millions of years) from the reference taxon. Within the same order Phthiraptera, these include, both a conspecific and congeneric taxon (1 and 2 respectively), a taxon in a different family (3), and two in different suborders (5 and 6). Finally, we selected a free-living taxon in closely related order (7) estimated to be 145 million years divergent from Phthiraptera. All age estimates were from Smith et al. (2011). Short Read Archive numbers are represented and those with (--) will be submitted upon acceptance.

TABLE 2: Results From aTRAM Assemblies

Genus	Species	Suborder	Family	Host	Species Code	SRA	[a]Total Genes assembled	[b]Passed Reciprocal Best Blast	[c]Total Genes >50%	[d]Total Nucleotides in Alignment
Antarctophthirus	*microchir*	Anoplura	Echinophthiriidae	*Otaria flavescens*	Aamic	SAMN05930899	1,095	1,056	943	856,295
Hoplopleura	*arboricola*	Anoplura	Hoplopleuridae	*Tamias amoenus*	Hbarb	SAMN05930902	1,077	1,044	932	879,297
Haematopinus	*eurysternus*	Anoplura	Haematopinidae	*Bos sp.*	Hieur	SAMN03360966	1,097	1,073	948	861,773
Linognathus	*spicatus*	Anoplura	Linognathidae	*Connochaetes taurinus*	Ltspi	SAMN03360967	1,091	1,066	945	889,413
Neohaematopinus	*pacificus*	Anoplura	Polyplacidae	*Tamias minimus*	Nepac	SAMN05930903	1,093	1,075	999	963,478
Pediculus	*humanus_1*	Anoplura	Pediculidae	*Homo sapien*	PdhumCA	SAMN05930905	1,084	1,059	908	808,011
Pediculus	*humanus_2*	Anoplura	Pediculidae	*Homo sapien*	PdhumHO	SAMN05930906	1,084	1,054	947	842,699
Pediculus	*schaeffi_1*	Anoplura	Pediculidae	*Pan troglodytes schweinfurthii*	PdschKE	SAMN05930907	1,103	1,083	1,039	988,430
Pediculus	*schaeffi_2*	Anoplura	Pediculidae	*Pan troglodytes schweinfurthii*	PdschUG	SAMN02438447	1,101	1,087	1,026	982,708
Pedicinus	*badii*	Anoplura	Pedicinidae	*Procolobus rufoimtratus*	Pnbad	SAMN03360969	1,085	1,062	861	791,014
Proechinophthirus	*fluctus*	Anoplura	Echinophthiriidae	*Callorhinus ursinus*	Prflu	SAMN03360968	1,087	1,050	867	769,201
Pthirus	*gorillae*	Anoplura	Pthiridae	*Gorilla beringei beringei*	Ptgor	SAMN05930908	1,106	1,077	1,037	983,085
Pthirus	*pubis*	Anoplura	Pthiridae	*Homo sapiens*	Ptpub	SAMN05930909	1,090	1,070	1,013	949,353
Brueelia	*antiqua*	Ischnocera	Ischnocera	*Catharus guttatus*	Brant	SAMN03360970	1,089	1,064	951	899,598
Bothriometopus	*macrocnemis*	Ischnocera	Ischnocera	*Chauna torquata*	Btmac	SAMN05930900	1,087	1,037	869	784,875

TABLE 2: Results from 1,107 targeted gene assembly with aTRAM across 13 sucking lice samples and 2 outgroups. The Genus, Species, Suborder, Family, Host and species code (unique for each taxon), are identified. The SRA (short read archive numbers are identified, and -- indicates those that will be submitted upon acceptance. [a]The number of genes that produced assembled contigs from aTRAM. [b]The number of

genes that survived the post processing exon extraction and reciprocal best blast test. [c]The number of genes with >50% of the exon sequences assembled relative to the reference *P. humanus humanus*. [d]The total nucleotides in the final concatenated alignment.