Phylogenomics from Low-coverage Whole-genome Sequencing

Article *in* Methods in Ecology and Evolution · January 2019

DOI: 10.1111/2041-210X.13145

CITATIONS
11

READS
1,230

7 authors, including:

Feng Zhang
Nanjing Agricultural University
86 PUBLICATIONS **698** CITATIONS
SEE PROFILE

Yinhuan Ding
Nanjing Agricultural University
15 PUBLICATIONS **103** CITATIONS
SEE PROFILE

Chao-Dong Zhu
Chinese Academy of Sciences
323 PUBLICATIONS **2,728** CITATIONS
SEE PROFILE

Xin Zhou
China Agricultural University
280 PUBLICATIONS **6,994** CITATIONS
SEE PROFILE

Received: 30 August 2018 | Accepted: 18 December 2018

DOI: 10.1111/2041-210X.13145

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Phylogenomics from low-coverage whole-genome sequencing

Feng Zhang[1,2,3] | Yinhuan Ding[1] | Chao-Dong Zhu[2,4] | Xin Zhou[5] | Michael C. Orr[2] | Stefan Scheu[3] | Yun-Xia Luan[6]

[1]Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing, P. R. China; [2]Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, P. R. China; [3]J. F. Blumenbach Institute of Zoology and Anthropology, University of Göttingen, Göttingen, Germany; [4]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, P. R. China; [5]Department of Entomology, China Agricultural University, Beijing, P. R. China and [6]Guangdong Provincial Key Laboratory of Insect Developmental Biology and Applied Technology, Institute of Insect Science and Technology, School of Life Sciences, South China Normal University, Guangzhou, P. R. China

Correspondence
Feng Zhang
Email: xtmtd.zf@gmail.com
and
Yun-Xia Luan
Email: yxluan@scnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31772491, 31772510 and 31625024; Key Laboratory of the Zoological Systematics and Evolution of the Chinese Academy of Sciences, Grant/Award Number: Y229YX5105

Handling Editor: Michael Matschiner

Abstract

1. Phylogenetic studies are increasingly reliant on next-generation sequencing. Transcriptomic and hybrid enrichment sequencing techniques remain the most prevalent methods for phylogenomic data collection due to their relatively low demands for computing powers and sequencing prices, compared to whole-genome sequencing (WGS). However, the transcriptome-based method is constrained by the availability of fresh materials and hybrid enrichment is limited by genomic resources necessary in probe designs, especially for non-model organisms.
2. We present a novel WGS-based pipeline for extracting essential phylogenomic markers through rapid de novo genome assembling from low-coverage genome data, employing a series of computationally efficient bioinformatic tools. We tested the pipeline on a Hexapoda dataset and a more focused Phthiraptera dataset (genome sizes 0.1–2 Gbp), and further investigated the effects of sequencing depth on target assembly success rate based on the raw data of six insect genomes (0.1–1 Gbp).
3. Each genome assembly was completed in 2–24 hr on desktop PCs. We extracted 872–1,615 near-universal single-copy orthologs (Benchmarking Universal Single-Copy Orthologs [BUSCOs]) per species. This method also enables the development of ultraconserved element (UCE) probe sets; we generated probes for Phthiraptera based on our WGS assemblies, containing 55,030 baits targeting 2,832 loci, from which we extracted 2,125–2,272 UCEs. Resulting phylogenetic trees all agreed with the currently accepted topologies, indicating that markers produced in our methods were valid for phylogenomic studies. We also showed that 10–20× sequencing coverage was sufficient to produce hundreds to thousands of targeted loci from BUSCO sets, and an even lower coverage (5×) was required for UCEs.
4. Our study demonstrates the feasibility of conducting phylogenomics from low-coverage WGS for a wide range of organisms without reference genomes. This

new approach has major advantages in data collection, particularly in reducing sequencing cost and computing consumption, while expanding loci choices.

KEYWORDS
desktop PC, genome assembly, hybrid enrichment, single-copy orthologs, ultraconserved elements

1 | INTRODUCTION

Advances in next-generation sequencing have greatly facilitated genome-scale data generation in the systematics community by enabling the collection of hundreds or thousands of loci for constructing phylogenies. Genomic partitioning (or "reduced representation") strategies, methods for enriching sequence libraries for selected genome regions (Turner, Ng, Nickerson, & Shendure, 2009), have dominated data collection approaches, given reduced computational burdens and costs compared to de novo whole-genome sequencing (WGS) (Jones & Good, 2016). Representative methods employed in deep phylogenetics include transcriptomic (RNA-seq; Wang, Gerstein, & Snyder, 2009) and hybrid enrichment sequencing (Bi et al., 2012; Briggs et al., 2009; Faircloth et al., 2012; Lemmon, Emme, & Lemmon, 2012). In recent years, these techniques have been successfully used to addressed a wide variety of questions in systematic and evolutionary biology (Fernández et al., 2018; Misof et al., 2014; Oakley, Wolfe, Lindgren, & Zaharoff, 2012; Prum et al., 2015; Young et al., 2016). Unfortunately, they have inherent practical limits (Lemmon & Lemmon, 2013). The transcriptomic approach requires a large quantity of high-quality RNA from fresh or carefully stored tissues (Cronn et al., 2012; McCormack, Hird, Zellmer, Carstens, & Brumfield, 2013). Hybrid enrichment techniques, such as anchored hybrid enrichment (AHE; Lemmon et al., 2012) and ultraconserved element (UCE) enrichment (Faircloth et al., 2012), have fewer limitations in material quality and quantity; but each specific group requires their own hybridization baits, and genomic resources are necessary to design these probe sets (Faircloth, 2017; Faircloth et al., 2012; Lemmon et al., 2012). This issue is exacerbated in small organisms, such as sucking lice or soil invertebrates, by few available genomic resources and a need for ample starting RNA/DNA. Perhaps the largest shortcoming of genome partitioning techniques is the narrow utility of the data generated, as these methods are rarely used outside of phylogenetic contexts (Allen et al., 2017).

The WGS has major advantages over genome partitioning methods in terms of material preparation, laboratory workload, diversity of targeted markers and future data utility. Until recently, the application of WGS in phylogenomics has been restricted by both high costs and computational challenges. With the emergence of new Illumina platforms (HiSeq X Ten, NovaSeq), sequencing costs have rapidly decreased, now as low as $10 per gigabase pairs (Novogene, China, April 1, 2018), thereby increasing economic feasibility of larger studies. Although genome assemblies are available, annotation and marker sorting are complicated and difficult processes, as seen in studies of birds (Jarvis et al., 2014). To address this issue, Allen, Huang, Cronk, and Johnson (2015) and Allen et al. (2017) developed an automated target restricted assembly method (aTRAM) which assembled targeted genes, rather than the entire genome, from WGS. Unfortunately, this approach still requires a relatively long time and a high computational memory because of BLAST tasks and assembly progress. Currently, all generalizable methods (RNA-based, AHE, aTRAM) which target protein-coding genes are hindered by laborious bioinformatic pipelines for orthology assignment and annotation. aTRAM may work for assembling multiple types of loci, such as UCEs or small circular genomes, but this has not been carefully tested (Allen et al., 2017).

Mining targeted loci directly from genome assemblies or WGS raw data is currently possible for some data types, including BUSCOs (Benchmarking Universal Single-Copy Orthologs; Waterhouse et al., 2018), UCEs (Faircloth, 2016), mitogenomes (Al-Nakeeb, Petersen, & Sicheritz-Pontén, 2017; Dierckxsens, Mardulyn, & Smits, 2017; Hahn, Bachmann, & Chevreux, 2013) and restriction site-associated DNA (Fan, Ives, & Surget-Groba, 2018). BUSCO assessments hold the potential to ameliorate the difficulties in orthology assignment by identifying near-universal single-copy orthologs (BUSCOs) (Waterhouse et al., 2018) based on the OrthoDB database (Zdobnov et al., 2017), a widely used resource for finding orthologs across diverse taxa. As such, BUSCOs have been applied to downstream phylogenetic inference in insects (Ioannidis et al., 2017), yeasts (Shen et al., 2016) and spiders (Fernández et al., 2018). However, assembling complete genomes from WGS data remains prohibitively computationally difficult, with even small- to medium-sized genomes typically requiring days for completion on dedicated servers. New, fast and memory-efficient de Bruijn graph (DBG) algorithms enable quicker genome assemblies on desktop computers (Chikhi, Limasset, & Medvedev, 2016; Chikhi & Rizk, 2013), but they have yet to be incorporated into WGS pipelines for targeting specific loci.

This study tests and improves the efficiency of mining popular phylogenomic markers (BUSCOs and UCEs) directly from low-coverage WGS data by rapidly assembling entire genomes for datasets across a wide range of taxa. All raw sequencing data and genome assemblies used here are retrieved from published studies and were of relatively low coverage (most below 30×). We integrate a series of fast and computationally efficient bioinformatic tools. Our pipeline applies read normalization (removing high-coverage reads) and the Minia3 assembler (Chikhi & Rizk, 2013) to greatly speed up genome assembly and extraction of extract single-copy genes. All analytical steps can be executed on desktop PCs in a relatively short

period of time (minutes to several hours for each step with the real datasets). We also test the assembly success rate of targeted loci at low to high depths of coverage on six insect genomes of various sizes (100–1,000 M). In doing so, we demonstrate the potential for this method to greatly improve current workflows for phylogenetic and other uses via WGS.

2 | MATERIALS AND METHODS

2.1 | Data generation

Our pipeline should be applicable to other similar datasets, even though we have demonstrated it only with Hexapoda datasets. The first dataset (A) includes 16 sucking lice species (Insecta: Phthiraptera) (Allen et al., 2017) and was selected for complete phylogenomic pipeline tests, including genome assembly, probe design, extraction of BUSCO/UCE loci and phylogenetic tree estimation (Table S1); initial assembled reads were subsampled to 4 G (11–34×, mean 24.19 ± 7.02, Table S5). A second real dataset B (14–47×, mean 26.38 ± 9.00, Table S6) of 21 species covering major hexapod lineages was used for BUSCO analyses, but not for UCE analyses because of great difficulties in designing a universal probe set for this highly divergent group. Two representative genomes, one small and one relatively large, were selected for each of five large orders (Hemiptera, Hymenoptera, Coleoptera, Lepidoptera and Diptera). These 21 species have genome sizes of 0.1–2 Gbp (Table S2). Six insect species with genome sizes ranging from 0.1 to 1 Gbp were used for assessments of assembly success rate (Table S3). Their UCE probe sets have been either published (Branstetter, Longino, Ward, & Faircloth, 2017; Faircloth, 2017) or were designed here (Phthiraptera).

The general workflow of our WGS phylogenomics pipeline has four main parts: data generation, genome assembly, loci extraction and phylogenetic inference (Figure 1). All assembly and data-mining analyses of the real datasets were executed in the CentOS 7 operating system on i7-7700 CPU (four cores/eight threads) and 16/32 G memory PCs, and others used a 24 cores/48 threads and 256 G memory server. Bioinformatic tools, custom scripts, and command details used in this study are given in the Supporting Information. Some steps can be omitted or replaced by other tools depending on the study's aims. Due to convenience and cost, raw sequencing data were typically generated on Illumina platforms (e.g., HiSeq 2500/X Ten, NovaSeq). Note that whole-genome amplification may

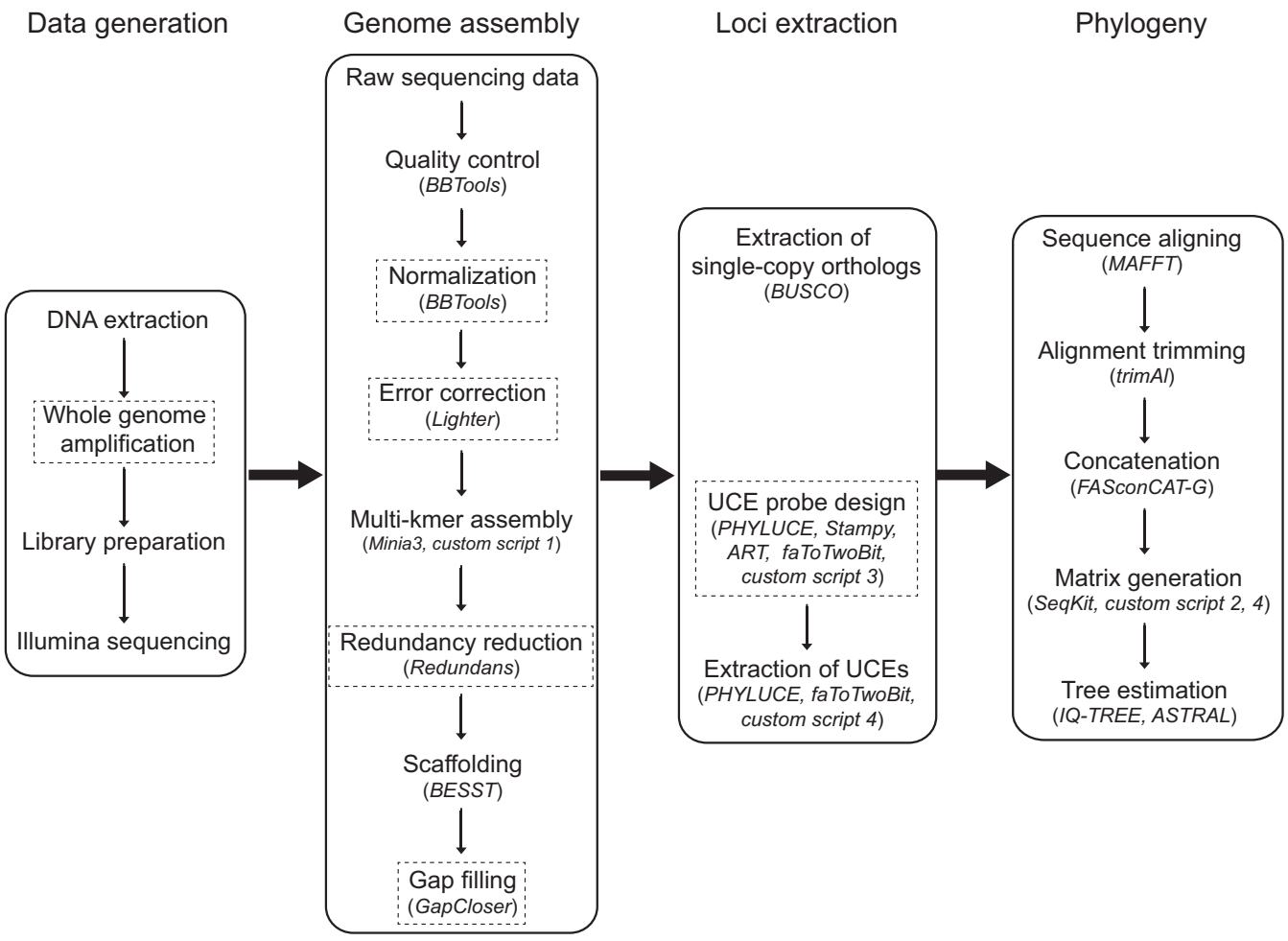

FIGURE 1 Flowchart of phylogenomics from whole-genome sequencing for assembling entire genomes. Bioinformatic tools used in each step are marked as italic. Dashed boxes indicate that these steps could be optionally omitted

be helpful for small organisms when the starting quantity of DNA does not meet the minimum criteria for WGS library preparation, although it may increase repeats and induce chimeras.

2.2 | Genome assembly

Raw sequencing data were downloaded and converted into gzipped fastq format with NCBI SRA TOOLKIT v2.9.0 (SRA Toolkit Development Team). Raw data of some species were subsampled at smaller sizes with reformat.sh (one of the BBTools suite, Bushnell, 2014). The resulting reads were compressed into clumps and duplicates were removed with clumpify.sh (BBTools). We used BBDUK.SH (BBTools) to perform quality trimming: both sides were trimmed to Q15 using the Phred algorithm, reads shorter than 15 bp or with more than 5 Ns were discarded, poly-A or poly-T tails of at least 10 bp were trimmed, and overlapping paired reads were corrected. To accelerate assembly and render difficult datasets tractable, we down-sampled reads over high-depth areas at an average depth of 10× by normalization using BBNORM.SH (BBTools). Sequencing errors were corrected with LIGHTER v1.1.1 (Song, Florea, & Langmead, 2014).

Genome contigs were assembled with multiple k-mer strategies using Minia3 and a custom script inspired by the GATB-Minia-Pipeline (https://github.com/GATB/gatb-minia-pipeline). K-mer values of 21, 41, 61, 81 were selected for read lengths around 100 bp, and 21, 41, 61, 81, 101, 121 for reads around 150 bp. Regions of high heterozygosity in diploid genomes are usually assembled as separate contigs once a pair of allelic sequences exceed a threshold of nucleotide diversity. These redundant contigs were removed using REDUNDANS v0.13c (Pryszcz & Gabaldón, 2016). Contig scaffolding and gap filling were performed with BESST v2.2.8 (Sahlin, Vezzi, Nystedt, Lundeberg, & Arvestad, 2014) and GAPCLOSER v1.12 in the SOAPdenovo2 suite (Luo et al., 2012) respectively. The input mapping file for scaffolding was generated with MINIMAP2 v2.9 (Li, 2018) and we then converted the mapping files into sorted, indexed BAM format using SAMTOOLS v1.7 (Li et al., 2009). A final genome assembly was generated for subsequent analyses.

2.3 | Single-copy orthologs

BUSCO v3.0.2 (Waterhouse et al., 2018) accepts both genomic and transcriptomic assemblies as inputs to generate complete, single-copy orthologs (BUSCOs) in "genome" mode using the predefined BUSCO sets. Assembly completeness is indicated by the ratio of complete and missing BUSCOs. A set of 1,658 loci was used for two hexapod datasets tested here. When the per cent of fragmented BUSCOs was >20% (331), BUSCO assessment was re-run by modifying the standard deviations (σ) of the mean BUSCO length to 2σ so that more BUSCOs were classified as "complete." Loci merging and aligning, alignment trimming and concatenating, and matrices generation and statistics were executed in a custom script integrating MAFFT v7.394 (Katoh & Standley, 2013), TRIMAL v1.4.1 (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009) and FASCONCAT-G v1.04 (Kück & Longo, 2014). Preliminary alignments were carried out using MAFFT with the L-INS-I strategy. Poorly aligned regions were automatically removed by the heuristic method automated1 with TRIMAL. Finally, we generated 50%–100% complete matrices. The completeness of a matrix represents the lowest ratio of taxa for all alignments. For example, a 100-taxa matrix of 75% completeness indicates that all alignments contain at least 75 taxa.

2.4 | UCE probe design and loci extraction

For groups lacking UCE probe sets, baits must be designed prior to loci identification. We followed Faircloth (2017) to design a bait set using 10 genome assemblies of Phthiraptera. The 10 species comprised three suborders (Anoplura, Ischnocera, Amblycera) and 10 families. One of them (*Pediculus humanus*) was selected as the base genome (accession GCA_000006295.1). The other nine exemplar genomes were assembled from WGS as in previous steps. *Osborniella crotophagae* (Amblycera) was treated as the "outgroup."

We simulated error-free, paired-end reads of 100 bp at a coverage of 2× with ART (Huang, Li, Myers, & Marth, 2012). The nine species' simulated reads were then aligned to the base genome using STAMPY v1.0.32 (Lunter & Goodson, 2011) with a substitution rate of 0.05, and unmapped reads were removed using SAMTOOLS. We merged overlapping or nearly overlapping intervals with bedtools (Quinlan & Hall, 2010). Putatively conserved intervals shared between nine exemplar species and the base genome were removed with phyluce_probe_strip_masked_loci_from_set (a python script within PHYLUCE v1.5.0, Faircloth, 2016). Shared, conserved loci between the base and exemplar species were determined with phyluce_probe_get_multi_merge_table. We extracted sequences with a length of 160 bp from the base genome that correspond to the loci we identified. A temporary bait set was designed targeting the above extracted loci with phyluce_probe_get_tiled_probes. Potentially problematic baits with >25% repeat content and GC content outside of the range of 30%–70% were removed. Duplicate baits were removed from this temporary bait set. We aligned the duplicate-free temporary baits against exemplar genomes with phyluce_probe_run_multiple_lastzs_sqlite to check if those loci could be located. FASTA data were then extracted from each of the exemplar genomes with phyluce_probe_slice_sequence_from_genomes. We determined those loci detected consistently across exemplar taxa with phyluce_probe_get_multi_fasta_table. We then designed the final bait set targeting those loci by tiling baits across each locus in each of 10 Phthiraptera genomes with phyluce_probe_get_tiled_probe_from_multiple_inputs. Putative duplicates were removed from the resulting bait set.

For UCE loci extraction, we aligned the probes to the genome sequences with phyluce_probe_run_multiple_lastzs_sqlite. FASTA sequences matching UCE loci were extracted from each genome by slicing 400 bp flanking region from both sides with phyluce_probe_slice_sequence_from_genomes. We then matched contigs to baits with phyluce_assembly_match_contigs_to_probes and phyluce_assembly_get_match_counts, and extracted all loci to a FASTA file with phyluce_assembly_get_fastas_from_match_counts. Similar to the

above BUSCO extraction, UCE aligning, trimming, concatenating and matrix generation and statistics were executed in a custom script.

2.5 | Phylogenetic analyses

The primary goals of this study were not phylogenetic, with trees used largely for checking the concordance of our methods with prior studies, and the analyses chosen reflect this. Some analyses, which may be also helpful for phylogenetic reconstructions, were omitted, including gene domain identification, sequence compositional heterogeneity, missing data distribution, locus screening, etc. We constructed the phylogenetic trees using maximum likelihood (ML) and coalescent-based species tree (ASTRAL) methods for both UCE and BUSCO matrices. Matrices of 100% (no missing taxa for all alignments) and 90% completeness (at most 10% missing taxa) were analysed as exemplars for Phthiraptera and Hexapoda respectively. Five matrices were generated: two BUSCO protein matrices (BUSCO_pro_A/B), two BUSCO nucleotide matrices (BUSCO_nuc_A/B) and one UCE nucleotide matrix (UCE_nuc_A). ML reconstructions were performed in IQ-TREE v1.6.3 (Nguyen, Schmidt, von Haeseler, & Minh, 2015) using partitioning schemes and substitution models that were automatically estimated with ModelFinder (Kalyaanamoorthy, Minh, Wong, von Haeseler, & Jermiin, 2017). Node supports were estimated using 1,000 ultrafast bootstrap (Hoang, Chernomor, von Haeseler, Minh, & Vinh, 2018) and 1,000 SH-aLRT replicates (Guindon et al., 2010). We restricted the procedure to a subset of substitution models with the options "-mset" (Hasegawa-Kishino-Yano [HKY] and generalised time-reversible [GTR] models for nucleotides, WAG and LG for proteins), and implemented the relaxed hierarchical clustering algorithm (Lanfear, Calcott, Kainer, Mayer, & Stamatakis, 2014) with the setting "-rcluster 10." For species tree estimation, gene trees were first estimated with IQ-TREE on individual gene alignments. Species trees were estimated from gene trees with ASTRAL-III v5.6.1 (Zhang, Rabiee, Sayyari, & Mirarab, 2018). Local branch supports on these species tree were estimated from quartet frequencies (Sayyari & Mirarab, 2016).

2.6 | Tests with varying sequencing coverage

To test the assembly success rate for our target loci, we performed the pipeline of genome assembly and loci extraction at depths of coverage of 1×, 5×, 10×, 20×, 30× with six insect species. Their genome sizes varied from 108 to 996 Mbp (Table S3). Raw input sequencing data were generated using reformat.sh on real data (Table S3).

3 | RESULTS

3.1 | Genome assembly

For dataset A, each Phthiraptera genome was assembled in 2–3 hr each on a 4-core/8-thread and 16 G memory PC. For dataset B, 21 hexapod genomes were assembled in 2–24 hr each on 4-core/8-thread and 16/32 G memory PCs. Basic statistics of assemblies and computational resource use are summarized in Tables S5 and S6. Number of scaffolds, maximum read length, N50 length, and GC content differed greatly among species.

3.2 | Extraction of single-copy orthologs

For dataset A, the detection rate of BUSCOs (complete and single-copy/duplicated + fragmented) reached 88.7%–98.2%. Among them, 1,162–1,615 (70.0%–97.4%) were classified as complete, single-copy BUSCOs with an average of 1,474 loci (88.9%) (Figure 3; Table S5). The final concatenated Phthiraptera matrices contained 475–1,627 BUSCOs of 211,055–688,462 amino acids or 630,964–2,078,753 nucleotide sites at a completeness level of 50%–100% (Figure 2).

For dataset B, the detection rate of all BUSCOs from WGS reached 65.3%–99.0% (Figure 2). Among them, 872–1,586 (1,310 ± 182.96) were complete, single-copy BUSCOs (Figure 4; Table S6). The proportion of fragmented and missing BUSCOs increased with genome size. Nine BUSCO analyses were re-run with modified length cutoffs because of the high proportion of fragmented BUSCOs. The

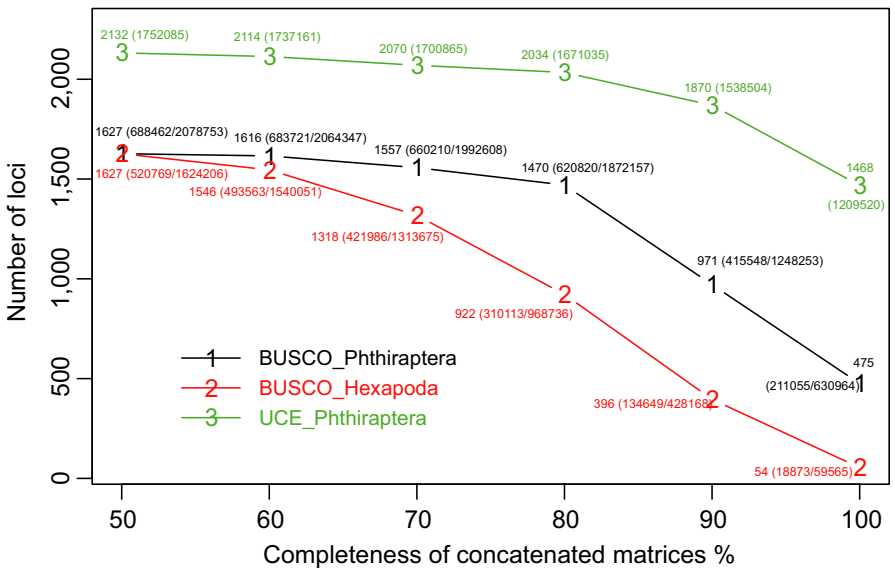

FIGURE 2 Number of loci and sites in the concatenated matrices of differing completeness for two real datasets. Both numbers (within parentheses) of amino acid and nucleotide sites are shown for complete single-copy Benchmarking Universal Single-Copy Orthologs (BUSCOs)

final concatenated hexapod matrix contained 54–1,627 BUSCOs of 18,873–520,769 amino acids or 59,565–1,624,206 nucleotide sites at a completeness of 50%–100% (Figure 2).

3.3 | Extraction of UCEs

Ultraconserved element analyses were performed on the Phthiraptera dataset. We simulated 1.83–3.13 M reads (mean 2.37 ± 0.43 M) from each exemplar genome assembly, and approximately 2.56–32.09% (mean 9.94 ± 8.64%) of these reads mapped to the base genome. We identified 5,356 conserved loci that were shared among *P. humanus* and all exemplar lineages. We designed 8,769 temporary baits from the base genome that target 4,743 conserved loci. A set of 2,882 conserved loci which were shared by *P. humanus* and the other nine species was selected for the final probe design. We then designed 56,001 baits targeting these 2,882 conserved loci based on all 10 taxa. Following removal of duplicates, the principle Phthiraptera bait set for UCE contained 55,030 baits targeting 2,832 conserved loci (named Phthiraptera-2.8Kv1).

Of the 2,832 targeted UCE loci, 75.0%–80.2% (2,125–2,272; average 2,205 = 77.9%) were extracted from 15 Phthiraptera species (Table S5). The final concatenated Phthiraptera matrix contained 1,468–2,132 UCEs of 1,209,520–1,752,085 nucleotide sites at a completeness of 50%–100% (Figure 2).

3.4 | Phylogenetic inference

The Phthiraptera matrix of 100% completeness was divided into 64, 76 and 169 partitions for matrices BUSCO_pro_A, BUSCO_nuc_A and UCE_nuc_A respectively. All ML and ASTRAL trees (Figures 3 and S1–S5) generated the same topology, congruent with phylogenies from the previous study (Allen et al., 2017), although supports at some nodes differed slightly. A node for (*Hoplopleura arboricola* + *Linognathus spicatus*) from UCE_nuc_A matrix had very low support values: 24.2/54 (Figure S2).

The Hexapoda matrix of 90% completeness was divided into 198 and 67 partitions for matrices BUSCO_pro_B and BUSCO_nuc_B respectively. Both ML and ASTRAL species trees (Figures 4 and S6–S8) generated topologies largely congruent with phylogenies from the published study (Misof et al., 2014). Only the positions of Thysanoptera and Psocodea were unstable among trees, indicated by lower node supports, possibly a result of inadequate sampling and crude phylogenetic analyses.

3.5 | Impacts of varying sequencing coverage

We tested the impact of sequencing coverage and genome size on the capture success rate of BUSCOs and UCEs using our pipeline. Basic statistics of genome assembly and loci extracted from data at the coverage

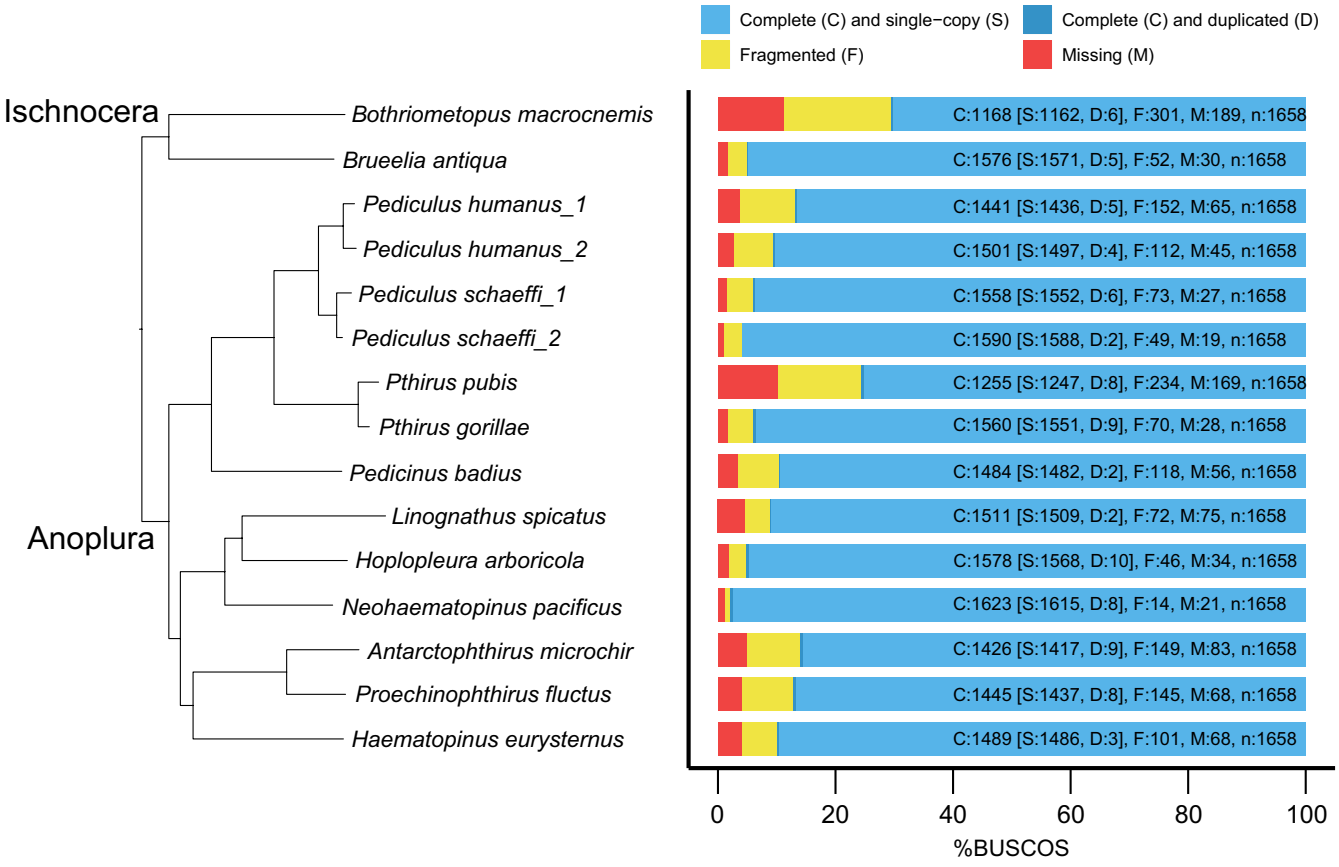

FIGURE 3 Maximum likelihood tree of Phthiraptera dataset based on concatenated Benchmarking Universal Single-Copy Orthologs (BUSCO) protein matrix of 100% completeness. Only node support values (SH-aLRT/UFBoot) below 100 are given in the tree. Right bar charts show BUSCO proportions classified as complete (C, blues), complete single-copy (S, light blue), complete duplicated (D, dark blue), fragmented (F, yellow) and missing (M, red)

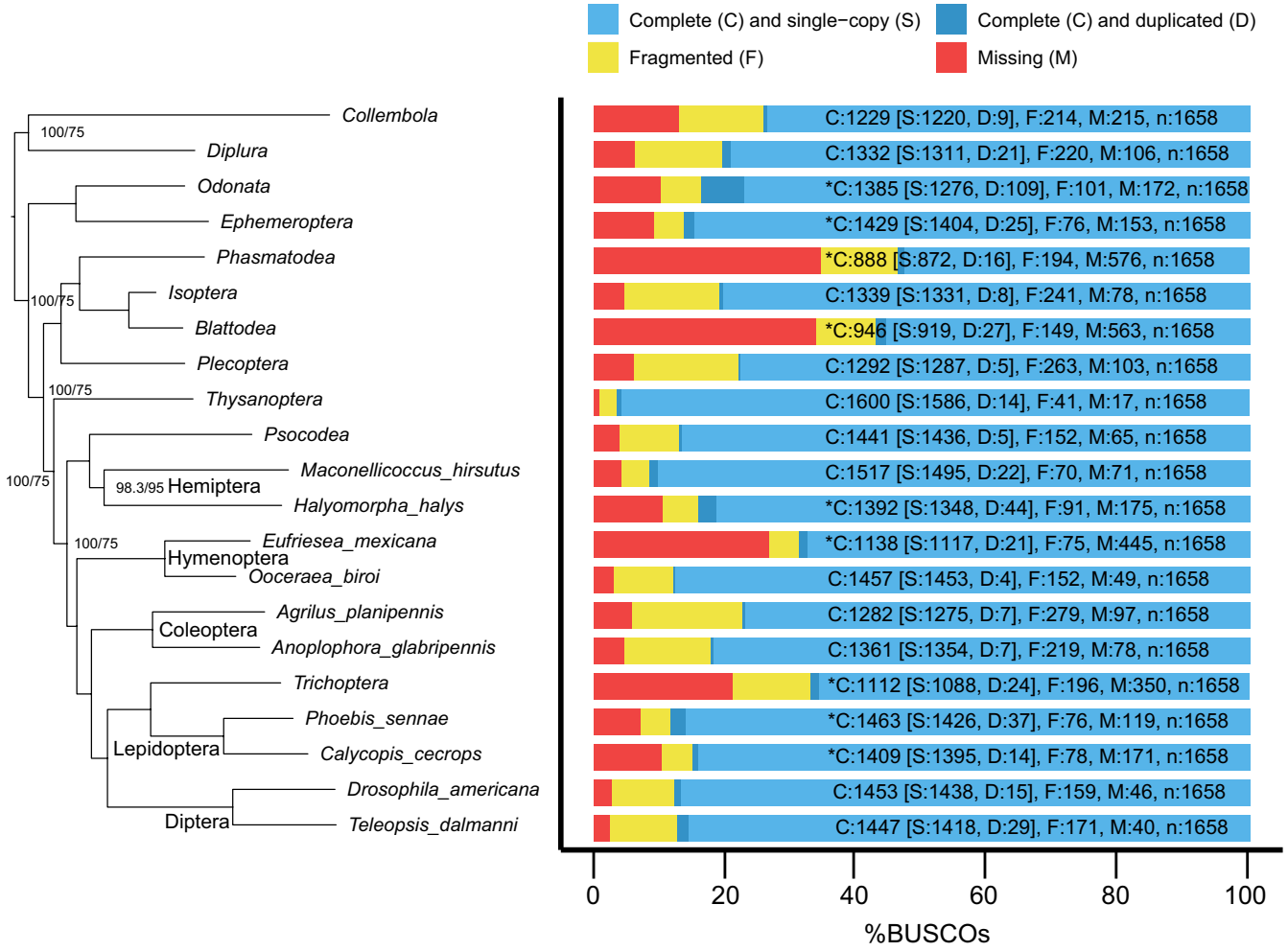

FIGURE 4 Maximum likelihood tree of Hexapoda dataset based on concatenated Benchmarking Universal Single-Copy Orthologs (BUSCO) nucleotide matrix of 90% completeness. Only node support values (SH-aLRT/UFBoot) below 100 are given in the tree. Right bar charts show BUSCO proportions classified as complete (C, blues), complete single-copy (S, light blue), complete duplicated (D, dark blue), fragmented (F, yellow) and missing (M, red). Asterisks represent results from BUSCO assessments using new length cut-offs

of 1×, 5×, 10×, 20×, 30×, and original reference assembly are summarized in Table S7. Most statistics reached convergence after 20×. The number of single-copy genes extracted from assemblies using BUSCO showed a similar trend, with an average of 3, 211, 914, 1,266, 1,330 and 1,543 for 1–30× and the reference genome respectively (Figure 5a; Table S7). UCEs required lower coverage, as the loci extracted from assemblies stabilized above just a coverage of 10× (Figure 5b; Table S7). The mean length of extracted UCEs increased significantly between coverages of 5–10×, and reached around 900 bp at the coverage of 20–30× (Figure 5c; Table S7).

4 | DISCUSSION

We successfully extracted hundreds to thousands of targeted loci from a single Illumina short-read library by assembling entire genomes using limited computational resources. Our study demonstrates the economic feasibility of phylogenomics using low-coverage WGS for a wide range of organisms with small-to-moderate (2,000 Mbp) genome sizes. Low coverage (10–20×) is feasible for BUSCOs and UCEs (Figure 5). A minimum coverage of 10× is recommended, consistent with the coverage requirement for aTRAM (Allen et al., 2017). There is little difference between 20× and 30× coverage in genome assemblies and number of BUSCOs. A coverage of 5× may be available for UCE data mining because the probe set was designed by simulating reads at a coverage of 2×. Notably, low coverage (<10×) usually generates relatively short length for UCEs. There are typically fewer BUSCOs extracted from larger genomes (>1 Gbp) than from smaller genomes (Figures 4 and 5a) because of the difficulties in assembling large genomes using short reads and a single library of small insert fragment sizes (see assembly statics in Tables S6 and S7). However, the number and length of UCEs extracted do not appear to be affected by genome size (Figure 5b,c).

4.1 | Merits of WGS

Generally, WGS outperforms transcriptomic and hybrid enrichment sequencing in terms of material preparation and laboratory protocol (Allen et al., 2017; Lemmon & Lemmon, 2013). WGS requires a lower

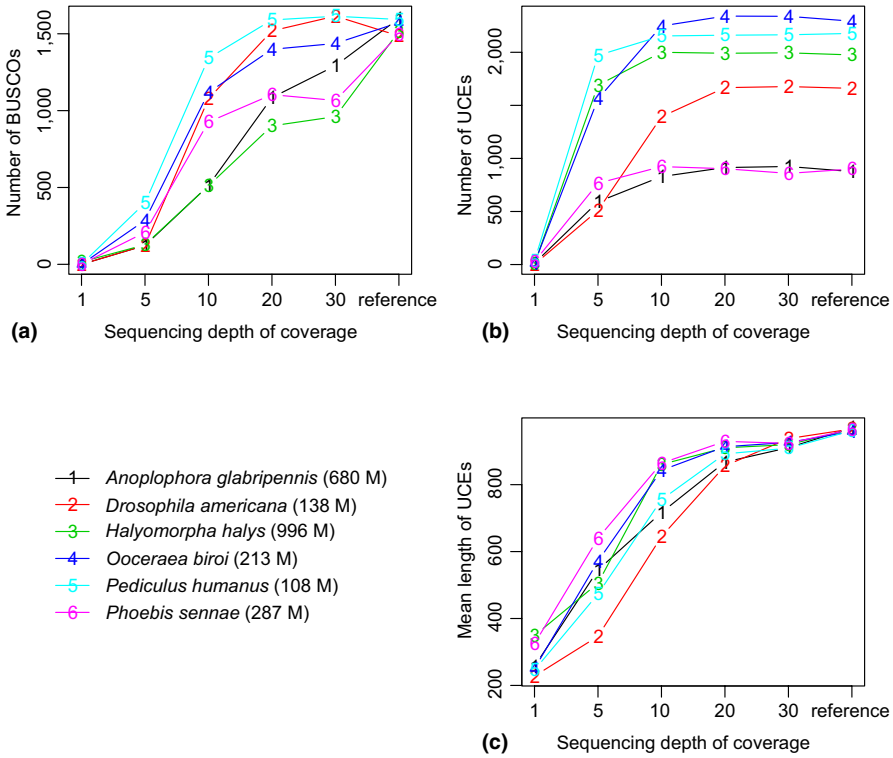

FIGURE 5 Impact on capture success rate of targeted loci number upon varying sequencing coverage for (a) complete single-copy genes (Benchmarking Universal Single-Copy Orthologs) and (b) ultraconserved elements (UCEs), as well as mean length of UCEs (c). Species are separated by different colours and numbers. "Reference" on the *x* axis represents the best genome assembly published for this species. Reference genome size for each species is shown following the species name

quantity of starting DNA, typically just 50–200 ng for an Illumina library. This method is possible even for very small organisms (size <1 mm, DNA < 10 ng) when augmented with whole-genome amplification using a multiple displacement amplification method (Dean et al., 2002; Lasken, 2009), comparable to DNA inputs in UCE protocol.

High WGS costs have previously hindered its application in phylogenomics, as systematists generally prefer to sequence as many taxa as possible. Now, the sequencing cost (library preparation of $30 and sequencing of $10/Gb) for a typical genome (size 0.1–1 G) with an average coverage of 15× is $45–180 on the HiSeq X Ten or NovaSeq platforms, a price not dissimilar to transcriptomic or hybrid enrichment sequencing (price from Novogene, China, April 1, 2018). Notably, the UCE approach via WGS may costs less because it requires lower coverage.

One of the biggest benefits of WGS approaches is that they are much more flexible in selection of loci type, thereby maximizing the use of the data collected (Allen et al., 2017). Entire genomes assembled from WGS have theoretical potential for all types of targeted regions, rather than for only one set, for example, mitochondrial genome assembly (Dierckxsens et al., 2017; Hahn et al., 2013). This procedure may even be useful for generating population genomics data, as several genomic population studies indicate that single nucleotide variants (SNVs) can be detected from WGS at low (3–12×) or extremely low (1×) coverage (Bizon et al., 2014; Rustagi et al., 2017). More genomic regions of interest should be tested in future studies, such as exons, introns, AHE loci and others.

4.2 | WGS analytical pipeline

Tremendous computational demands of time and resources, as well as high cost, are additional major challenges for WGS projects. A complete genome assembly of a eukaryotic organism usually requires several days or weeks on a dedicated server or cluster, which is not feasible for many small laboratories. With our method, using real datasets, we executed genome assemblies and extracted BUSCOs in 2–24 hr each on desktop PCs (Tables S5 and S6). This procedure will take more time with larger genomes but it will still drastically improve upon current time frames.

The proposed novel pipeline employs a series of computationally efficient bioinformatic tools, enabling execution of all analyses on a desktop PC in minimal time. Its workflow is flexible, custom-made and some steps may be omitted or replaced by other tools depending on study aims and genome features. Many analyses requiring additional configuration files and original manuals are simplified with our custom scripts, which automatically generate them. When the required tools are ready, the main steps can be implemented using a single bash script. Compared to aTRAM, another phylogenomic approach from low-coverage WGS, our pipeline can be executed in a shorter time on a desktop computer.

Our optimization steps primarily focus on genome assembly, particularly read normalization and low-consumption assembly. De novo assemblies that rely on DBG usually consume a lot of memory. To speed up assembly and reduce computational burdens, the assembler Minia3 uses a novel data structure to construct compacted graphs (Chikhi et al., 2016). K-mer–based normalization can also dramatically accelerate assembly by removing redundant short reads with little to no change in the overall assembly quality (Brown, Howe, Zhang, Pyrkosz, & Brom, 2012). With the tool BBNorm, regions below the target coverage will be retained and those with coverage above the target will be reduced to the target, further simplifying analyses.

Streamlining is also possible for post-assembly analyses (Figure 1). Heterozygous regions containing one or more heterozygous sites can be problematic for downstream analyses, such as paralog identification and collinearity analyses (Pryszcz & Gabaldón, 2016). Most phylogenetic samples are collected from wild populations rather than inbreeding strains and thus may have a relatively high rate of heterozygosity. Removal of these redundant contigs is often useful in generating more complete, single-copy than duplicated BUSCOs and reducing the subsequent computational burden (Table S7). The process is further simplified with BUSCO, as the use of single-copy orthologs avoids the laborious analyses of orthology assignment and annotation while still assessing genome completeness (Waterhouse et al., 2018). The present version, BUSCO v3, includes universal gene sets for most biological lineages. UCE probe sets are still lacking in most metazoan groups, but our method also enables quicker and more efficient design of probe sets via WGS.

4.3 | Current limits of low-coverage WGS approach

Undoubtedly, higher quality genome assemblies will improve the extraction of targeted loci. Low-coverage, short-read sequencing and a single library of short insert size certainly perform worse for large genomes, as shown in this study (Figure 4; Table S6). This is because short reads increase the complexity of the assembly algorithms, especially for repeated or heterozygous regions (Miller, Koren, & Sutton, 2010), although high coverage and longer library insert sizes can overcome some issues (Wetzel, Kingsford, & Pop, 2011). Low N50 and a high proportion of fragmented BUSCOs indicated inferior assembly contiguity. At the current stage, there are few methods for amelioration except for performing multiple library strategies or using new sequencing platforms (PacBio/Nanopore), which are more expensive. In spite of poor assembly quality for large genomes, BUSCO assessments can provide hundreds of fragmented and complete orthologs for further phylogenomic analyses. Contiguity may also be improved by relaxing the length cut-off of "complete" BUSCOs so that more loci may be used. Species-specific training parameters for Augustus prediction (Keller, Kollmar, Stanke, & Waack, 2011) can also improve BUSCO performance. In contrast to BUSCOs, UCEs and SNVs compatible with lower coverage have fewer limits in terms of target capture. In addition, phylogenetic signal and the efficiency of predefined reference genes are rarely tested, important procedures absent in most studies. Therefore, the construction of lineage-specific datasets should substantially improve both the mining accuracy of linage-specific, single-copy orthologs and resultant phylogenetic estimates.

Although our pipeline performed well across insect orders, its efficacy must be tested in more organisms, as well as more genomic regions of interest. With the development of new sequencing and assembly techniques, we believe that WGS will increasingly dominate phylogenetic studies.

ACKNOWLEDGEMENTS

All authors declare that there is no conflict of interest. We thank the editor and reviewers for their valuable input. This work was supported by the National Natural Science Foundation of China (31772491, 31772510); and a grant from the Key Laboratory of the Zoological Systematics and Evolution of the Chinese Academy of Sciences (Y229YX5105). Chao-Dong Zhu acknowledges the supports of the National Science Fund for Distinguished Young Scholars (31625024).

AUTHORS' CONTRIBUTIONS

F.Z. and Y. -X. L. conceived the ideas, designed methodology and led the writing of the manuscript; F.Z. and Y.D. collected and analysed the data; C.Z., X.Z., M.C.O. and S.S revised the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

Raw sequencing data were downloaded from GenBank (accession numbers SRR5088465, SRR5088468, SRR5308123, SRR5308129, SRR5088469, SRR5088471, SRR5088472, SRR5088473, SRR1182279,SRR5308136,SRR5308138,SRR5088474,SRR5088475,SRR5308112,SRR5088466,SRR5088470,SRR5626543,SRR1727714,SRR3547158, SRR3262630, SRR3262386, SRR5192509, SRR5630825, SRR863588, SRR1300141, SRR1300143, ERR1189167, SRR1298381, SRR5088472, SRR1192097, SRR1945067, SRR1174017, SRR947086, SRR947087, SRR3091597, SRR3397505, SRR5750546, SRR941726, SRR941727, SRR1298382, ERR957820, ERR957821). The list of bioinformatic tools and customs scripts is available at https://github.com/xtmtd/PLWS (https://doi.org/10.5281/zenodo.2492625). UCE probe sets for Phthiraptera and all data files involving phylogenetic analyses are available from the figshare (https://figshare.com/projects/Phylogenomics_from_Low-coverage_Whole-genome_Sequencing/56942).

ORCID

Feng Zhang https://orcid.org/0000-0002-1371-266X
Chao-Dong Zhu https://orcid.org/0000-0002-9347-3178
Xin Zhou https://orcid.org/0000-0002-1407-7952
Michael C. Orr https://orcid.org/0000-0002-9096-3008
Yun-Xia Luan https://orcid.org/0000-0003-3573-7144

REFERENCES

Allen, J. M., Boyd, B., Nguyen, N. P., Vachaspati, P., Warnow, T., Huang, D. I., … Johnson, K. P. (2017). Phylogenomics from whole genome sequences using aTRAM. *Systematic Biology*, *66*, 786–798. https://doi.org/10.1093/sysbio/syw105

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. https://doi.org/10.1186/s12859-015-0515-2

Al-Nakeeb, K., Petersen, T. N., & Sicheritz-Pontén, T. (2017). Norgal: Extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, *18*, 510. https://doi.org/10.1186/s12859-017-1927-y

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*, 403. https://doi.org/10.1186/1471-2164-13-403

Bizon, C., Spiegel, M., Chasse, S. A., Gizer, I. R., Li, Y., Malc, E. P., … Wilhelmsen, K. C. (2014). Variant calling in low-coverage whole genome sequencing of a Native American population sample. *BMC Genomics*, *15*, 85. https://doi.org/10.1186/1471-2164-15-85

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*, 768–776. https://doi.org/10.1111/2041-210x.12742

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., … Paabo, S. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, *325*, 318–321. https://doi.org/10.1126/science.1174462

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., & Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv*, 1203.4802.

Bushnell, B. (2014). *BBtools*. Retrieved from https://sourceforge.net/projects/bbmap/

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chikhi, R., Limasset, A., & Medvedev, P. (2016). Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, *32*, i201–i208. https://doi.org/10.1093/bioinformatics/btw279

Chikhi, R., & Rizk, G. (2013). Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for Molecular Biology*, *8*, 22. https://doi.org/10.1186/1748-7188-8-22

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., & Udall, J. (2012). Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, *99*, 291–311. https://doi.org/10.3732/ajb.1100356

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., … Lasken, R. S. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 5261–5266. https://doi.org/10.1073/pnas.082089499

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*, e18. https://doi.org/10.1093/nar/gkw955

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*, 786–788. https://doi.org/10.1093/bioinformatics/btv646

Faircloth, B. C. (2017). Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution*, *8*, 1103–1112. https://doi.org/10.1111/2041-210x.12754

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. *Systematic Biology*, *61*, 717–726. https://doi.org/10.1093/sysbio/sys004

Fan, H., Ives, A. R., & Surget-Groba, Y. (2018). Reconstructing phylogeny from reduced-representation genome sequencing data without assembly or alignment. *Molecular Ecology Resources*, *18*, 1482–1491. https://doi.org/10.1111/1755-0998.12921

Fernández, R., Kallal, R. J., Dimitrov, D., Ballesteros, J. A., Arnedo, M. A., Giribet, G., & Hormiga, G. (2018). Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Current Biology*, *28*, 1489–1497. e5. https://doi.org/10.1016/j.cub.2018.06.018

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*, 307–321. https://doi.org/10.1093/sysbio/syq010

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—A baiting and iterative mapping approach. *Nucleic Acids Research*, *41*, e129. https://doi.org/10.1093/nar/gkt371

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*, 518–522. https://doi.org/10.1093/molbev/msx281

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*, 593–594. https://doi.org/10.1093/bioinformatics/btr708

Ioannidis, P., Simão, F. A., Waterhouse, R. M., Manni, M., Seppey, M., Robertson, H. M., … Zdobnov, E. M. (2017). Genomic features of the damselfly *Calopteryx splendens* representing a sister clade to most insect orders. *Genome Biology and Evolution*, *9*, 415–430. https://doi.org/10.1093/gbe/evx006

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., … Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, *346*, 1320–1331. https://doi.org/10.1126/science.aab1062

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*, 185–202. https://doi.org/10.1111/mec.13304

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*, 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, *27*, 757–763. https://doi.org/10.1093/bioinformatics/btr010

Kück, P., & Longo, G. C. (2014). FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*, *11*, 81. https://doi.org/10.1186/s12983-014-0081-x

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology*, *14*, 82. https://doi.org/10.1186/1471-2148-14-82

Lasken, R. S. (2009). Genomic DNA amplification by the multiple displacement amplification (MDA) method. *Biochemical Society Transactions*, *37*, 450. https://doi.org/10.1042/bst0370450

Lemmon, A. R., Emme, S., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*, 727–744. https://doi.org/10.1093/sysbio/sys049

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009).

The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*, 936–939. https://doi.org/10.1101/gr.111120.110

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*, 18. https://doi.org/10.1186/2047-217x-1-18

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, *66*, 526–538. https://doi.org/10.1016/j.ympev.2011.12.007

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*, 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., … Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*, 763–767. https://doi.org/10.1126/science.1257570

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Molecular Biology and Evolution*, *32*, 268–274. https://doi.org/10.1093/molbev/msu300

Oakley, T. H., Wolfe, J. M., Lindgren, A. R., & Zaharoff, A. K. (2012). Phylotranscriptomics to bring the understudied into the fold: Monophyletic ostracoda, fossil placement and pancrustacean phylogeny. *Molecular Biology and Evolution*, *30*, 215–233. https://doi.org/10.1093/molbev/mss216

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. https://doi.org/10.1038/nature19417

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*, e113. https://doi.org/10.1093/nar/gkw294

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rustagi, N., Zhou, A., Watkins, W. S., Gedvilaite, E., Wang, S., Ramesh, N., … Xing, J. (2017). Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. *BMC Genomics*, *18*, 396. https://doi.org/10.1186/s12864-017-3767-6

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., & Arvestad, L. (2014). BESST–efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, *15*, 281. https://doi.org/10.1186/1471-2105-15-281

Sayyari, E., & Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, *33*, 1654–1668. https://doi.org/10.1093/molbev/msw079

Shen, X. X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., & Rokas, A. (2016). Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. *G3: Genes, Genomes, Genetics*, *6*, 3927–3939. https://doi.org/10.1534/g3.116.034744

Song, L., Florea, L., & Langmead, B. (2014). Lighter: Fast and memory-efficient sequencing error correction without counting. *Genome Biology*, *15*, 509. https://doi.org/10.1186/s13059-014-0509-9

Turner, E. H., Ng, S. B., Nickerson, D. A., & Shendure, J. (2009). Methods for genomic partitioning. *Annual Review of Genomics and Human Genetics*, *10*, 263–284. https://doi.org/10.1146/annurev-genom-082908-150112

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*, 57–63. https://doi.org/10.1038/nrg2484

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., … Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*, 543–548. https://doi.org/10.1093/molbev/msx319

Wetzel, J., Kingsford, C., & Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics*, *12*, 95. https://doi.org/10.1186/1471-2105-12-95

Young, A. D., Lemmon, A. R., Skevington, J. H., Mengual, X., Stahls, G., Reemer, M., … Wiegmann, B. M. (2016). Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC Evolutionary Biology*, *16*, 143. https://doi.org/10.1186/s12862-016-0714-0

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., … Kriventseva, E. V. (2017). OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, *45*, D744–D749. https://doi.org/10.1093/nar/gkw1119

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*, 153. https://doi.org/10.1186/s12859-018-2129-y

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zhang F, Ding Y, Zhu C-D, et al. Phylogenomics from low-coverage whole-genome sequencing. *Methods Ecol Evol*. 2019;00:1–11. https://doi.org/10.1111/2041-210X.13145