Chromosome-Level Genome Assembly of the Heptageniid Mayfly *Parafronurus youi* (Ephemeroptera), and Its Annotation

Ran Li [1,*], Ze-Kai Wang [1], Dong-Kai Liu [1], Ying-Xue Zhang [1], Xiao-Yu Li [1], Hai-Xin Li [1]

[1]School of Life Sciences, Qufu Normal University, 273165 Qufu, China

*Corresponding author: E-mail: li471329014@163.com

Abstract

As a group of winged insects (Pterygota) retaining many primitive characteristics, genomic research on mayflies remains highly limited, posing challenges to the study of their origin and evolution. In this study, we present the first chromosome-level genome assembly of the Chinese endemic mayfly *Parafronurus youi* utilizing Illumina short-read, PacBio long-read, and Hi-C sequencing technologies. The high-quality genome is 412.90 Mb in size with 99.07% of the sequences anchored to 11 chromosomes (ranging from 24.88 to 45.89 Mb). Genome annotation predicted 15,647 protein-coding genes with an average length of 9,934.7 bp, of which 85.9% were functionally annotated in the UniProtKB database. Repetitive elements accounted for 32.83% of the genome, including 27.33% transposable elements and 4.07% simple repeats. This study not only enriches genomic resources for mayflies but also establishes a foundation for investigating molecular mechanisms underlying ecological adaptation and evolutionary traits, contributing to the conservation of freshwater ecosystems.

Key words: Ephemeroptera, *Parafronurus youi*, genome assembly, gene annotation, mayfly

Significance

By providing the first chromosome-level genome assembly for *Parafronurus youi*, the study bridges a significant gap in genomic resources for Ephemeroptera, facilitating deeper insights into their unique evolutionary traits, ecological roles, and environmental adaptations. Moreover, the high-quality genomic data support the exploration of genetic mechanisms underlying adaptation to freshwater habitats and their utility as bioindicators.

Introduction

The order Ephemeroptera, commonly known as mayflies, is one of the most ancient groups of winged insects (Pterygota), with a fossil record dating back to the Carboniferous period around 300 million years ago (Bauernfeind and Soldάn 2012). Mayflies retain numerous primitive and unique characteristics, including their distinctive prometamorphosis development pattern and uniquely evolved wing venation (Sartori and Brittain 2015; Kamsoi et al. 2021). These traits make them an essential group for studying the origin and evolution of insect wings. Currently, Ephemeroptera comprises 42 families and 450 genera, with approximately 4,000 species described worldwide (Jacobus et al. 2021). Although the winged adult stage is brief and primarily focused on reproduction, mayflies spend the majority of their lives as aquatic nymphs, where they play critical roles in nutrient cycling and food web dynamics. Their sensitivity to water quality also makes them valuable bioindicators, offering insights into freshwater ecosystem health and serving as important tools for environmental monitoring and conservation efforts (Jacobus et al. 2019).

Over the past few decades, researches on mayflies have primarily centered on taxonomy, ecology, behavior, and biogeography (Zheng et al. 2024; El Yaagoubi et al. 2024; Mayorga et al. 2024; Srinivasan et al. 2024). Molecular phylogenetic studies have mainly relied on traditional morphological traits, mitochondrial genomes, or a limited number of nuclear genes (Khudhur and Shekha 2020; Li et al. 2020; Wakimura et al. 2024). Genomic research on mayflies remains underdeveloped compared to other insect groups, with data available for only seven species, three of which have high-resolution chromosome-level assemblies (*Cloeon dipterum*, GCF_949628265.1; *Siphlonurus alternatus*, GCA_949825025.1; *Ecdyonurus torrentis*, GCA_949318235.1) (Farr et al. 2023a; Farr et al. 2023b; Farr et al. 2023c). However, only *C. dipterum* has undergone detailed annotation, leaving key questions about genomic structure, function, and evolutionary mechanisms of Ephemeroptera unanswered. The lack of high-quality reference genomes also limits the identification of genes related to environmental adaptations and evolutionary traits. As a basal group within Pterygota, the chromosomal evolution of mayflies remains largely unexplored, emphasizing the need for robust genomic resources.

To address this gap, we present the first chromosome-level genome sequencing, assembly, and annotation of *Parafronurus youi*, a mayfly species endemic to China that inhabits clean, fast-flowing mountain streams and rivers. This study provides a high-resolution reference genome, offering critical insights into the genomic structure and evolution of *P. youi*. By expanding genomic resources for Ephemeroptera, this research establishes a foundation for future studies in evolutionary biology, functional genomics, and phylogenomics, while contributing to efforts to protect and sustain freshwater ecosystems.

Results and Discussion

Genome Estimation

To preliminarily assess the genomic characteristics of *P. youi*, we generated 41.84 Gb of Illumina sequencing data (Table S1). Analysis of clean reads through k-mer distribution estimated the genome size to range between 386.73 Mb and 387.50 Mb (k-mers peaking at 21), with the final estimated genome size determined as 387.503 Mb (supplementary fig. S1 and table S2). Further characterization of genome properties estimated a heterozygosity rate of 0.62% and a repetitive sequence content of 24.27% (supplementary table S2). These results provide critical baseline data for understanding the genomic organization of *P. youi* and set the stage for downstream analyses.

Genome Assembly and Assessment

The genome was assembled using a total of 60.18 Gb of PacBio long-read sequencing data, yielding ~155.1× coverage of the estimated genome size. The PacBio subreads had a mean length of 26.13 kb, with an N50 of 22.49 kb (supplementary table S1). The initial assembly produced a genome size of 557.58 Mb with a contig N50 of 0.45 Mb (supplementary table S3), which exceeded the estimated genome size due to the inclusion of redundant regions. To refine the assembly, one round of polishing with long reads and two rounds with Illumina short reads were performed, followed by the removal of redundant sequences. The resulting draft assembly was reduced to 413.32 Mb, with an N50 of 0.59 Mb, comprising 1,099 contigs (supplementary table S3). To achieve a chromosome-level assembly, 42.72 Gb of Hi-C sequencing data (284,784,594 reads) was utilized to anchor and orient contigs into pseudo-chromosomes. This process successfully mapped 99.07% of the total genome assembly to 11 chromosomes, resulting in a final assembly size of 412.90 Mb, consisting of 219 scaffolds with an N50 of 41.95 Mb (Fig. 1a, Table 1, and supplementary table S3). The chromosome lengths ranged from 24.88 to 45.89 Mb (supplementary table S4). Results showed that the size of the assembled mayfly genome was slightly larger than the estimated size. A possible reason for this discrepancy is that k-mer-based estimation typically relies on the frequency distribution of unique k-mers, which may underestimate the contribution of highly repetitive sequences. Notably, our genome was significantly larger than that of *C. dipterum* (190.1 Mb), and comparable in size to the two other chromosome-level genomes, *S. alternatus* (455.8 Mb) and *E. torrentis* (503.2 Mb), all of which shared the same chromosome number of 11 (Farr et al. 2023a; Farr et al. 2023b).

The quality and completeness of the genome assembly were validated using BUSCO analysis, which identified 96.7% of single-copy orthologs as complete (94.1% single-copy and 2.6% duplicated), with 1.0% fragmented and 2.3% missing (supplementary table S3). Furthermore, mapping rates of 90.47%, 95.75%, and 98.81% were achieved for Illumina, RNA-seq, and PacBio data, respectively, confirming the high accuracy and integrity of the assembled genome. These results demonstrate that the genome of *P. youi* provides a robust resource for further biological and ecological studies.

Genome Annotation

Repetitive elements accounted for 32.83% of the *P. youi* genome, encompassing 135.57 Mb. Among these, transposable elements (TEs) represented the majority (27.33%), with the remainder comprising simple repeats (4.07%), low-complexity regions (0.67%), small RNAs (0.74%), and satellite sequences (0.17%) (supplementary table S5). A significant portion of the TEs (20.34%) remained unclassified due to a lack of homologs in known databases. DNA transposons (3.31%) and LINEs (2.37%) were the most abundant classified elements (Fig. 1b, supplementary table S5).

Non-coding RNA (ncRNA) annotation revealed diverse RNA types, including 261 tRNAs (representing 21 isotypes, Supres lacking), 47 small nuclear RNAs (snRNAs), 433 ribosomal RNAs (rRNAs), 72 micro RNAs (miRNAs), 421 ribozyme RNAs, two long non-coding RNAs (lncRNAs), one leader RNA and 142 other types of ncRNAs (supplementary table S6). Of the snRNAs, spliceosomal RNAs dominated, with 17 classified into subtypes such as U1, U2, U4, U5, U6, and U11. Additionally, 25 C/D box snoRNAs, one SCARNA8, and one SNORA16 were identified.

After masking repetitive sequences and ncRNAs, a total of 15,647 protein-coding genes (PCGs) were predicted, with an average gene length of 9,934.7 bp and an average coding sequence (CDS) length of 194.1 bp. Each gene contained an average of 11.2 exons, with an average exon length of 270.9 bp and an intron length of 964.9 bp. BUSCO analysis of the predicted genes identified 97.6% of insect single-copy orthologs (insecta_odb10 database) as complete, with 75.5% being single-copy, 22.1% multi-copy, further confirming the accuracy of the annotation (Table 1). Functional annotation of PCGs revealed that 85.9% (13,441 genes) were successfully matched to the UniProtKB database. Additional classifications using InterProScan and EggNOG identified 10,776 genes associated with GO terms and 10,228 genes mapped to KEGG pathways. Furthermore, the analysis provided extensive annotations, including 8,189 KO terms, 2,820 enzyme codes, 11,491 Reactome pathways, and 12,100 COG categories. In total, 13,975 genes were annotated by at least one database. Comparison with *C. dipterum* highlighted distinct genomic features of *P. youi*, including a higher number of predicted protein-coding genes (15,647 vs. 13,282) and more extensive functional annotations. These results emphasize the value of the *P. youi* genome in advancing our understanding of mayfly biology and evolution.

Methods

Sample collection and DNA/RNA Extraction

Specimens of *P. youi* were captured from Zijin Mountain, Nanjing, China. To preserve integrity, samples were immediately frozen in liquid nitrogen and transported to the laboratory. Genomic DNA was extracted using the Blood & Cell Culture DNA Mini Kit (Qiagen), while RNA was isolated using the TRIzol Total RNA Isolation Kit (Rio et al. 2010). DNA and RNA quality were assessed with a Qubit 3.0 Fluorometer and 1% agarose gel electrophoresis to ensure high-quality input material.

Library Preparation and Sequencing

High-quality DNA was used to construct single-molecule real-time (SMRT) sequencing libraries with an insert size of 30 kb using the SMRTbell$^{\text{TM}}$ Template Prep Kit 2.0 (PacBio). Long-read sequencing was conducted on the PacBio Sequel II platform. For short-read sequencing, paired-end libraries (350 bp insert size) were prepared using the TruSeq DNA PCR-free kit (Illumina) and sequenced on the NovaSeq 6000 platform. Transcriptome sequencing libraries were constructed using the TruSeq RNA Sample Prep Kit and sequenced with PE150 reads. Hi-C libraries were prepared using standard proximity ligation protocols to capture chromatin conformation, followed by sequencing on the Illumina NovaSeq 6000 platform. Quality control of raw Illumina reads was performed using BBTools v38.82, ensuring clean and high-quality data (Bushnell 2014). PacBio long subreads were directly produced by the sequencing equipment (Sequel II system). Hi-C sequencing data underwent quality control using Juicer v1.6.2 (Durand et al. 2016b).

Genome Size Estimation

Filtered Illumina data was used for genome survey analysis to estimate genome characterize with the k-mer length of 21. The number and distribution of k-mer were calculated using "khist.sh" (BBTools v38.82). Genome characteristics were further visualized and analyzed using GenomeScope v2.0 (Vurture et al. 2017; Marcais and Kingsford 2011). The heterozygosity ratio was calculated based on the heterozygous peak value.

Genome assembly and Assessment

Filtered long reads were subject to *de novo* assembly using NextDenovo v. 2.4.0 (Hu et al. 2024b). The primary genome assembly was polished using NextPolish 1.4.12 with one round based on long reads and two rounds of short reads (Hu et al. 2024a). The heterogeneous regions (e.g. haplotigs and overlaps) was removed using Purge dups v1.0.1 (Roach et al. 2018). Chromosome-level genome assembly was performed using the 3D-DNA v180922 pipeline, and manual correction of assembly errors, contig orientations, and chromosomal boundaries was conducted using JuiceBox v1.11.08 (Durand et al. 2016a). The final assembly underwent contamination detection to ensure high-quality results.

Genome completeness was evaluated using BUSCO v5.4.4 against the insecta_odb10 database (n = 1,367) (Waterhouse et al. 2018). Sequencing reads were mapped to the final assembly using Minimap2 v2.23 to validate genome integrity. High mapping rates for Illumina, RNA-seq, and PacBio data supported the assembly’s robustness and accuracy

(Li et al. 2018).

Annotation of Repeats and Genes

Repetitive elements were detected using a combination of *ab initio* and homology-based prediction methods. RepeatModeler v2.0.2 was employed to identify the repetitive elements, and a custom library combining Dfam v3.1 and RepBase v20181026 were constructed (Flynn et al. 2020; Hubley et al. 2016； Bao et al. 2015). RepeatMasker v4.1.2 was employed to mask repeats against this library (Smit et al. 2015). Non-coding RNAs were annotated using tRNAscan-SE v2.0.7 and Infernal v1.1.3, retaining only high-confidence predictions (Nawrocki and Eddy 2013; Chan et al. 2019).

Protein-coding genes (PCGs) were predicted using MAKER v3.01.03, an integrative pipeline combining *ab initio*, transcriptome-based, and homology-based methods (Hoff and Yandell 2011). The *ab initio* predictions were carried out using BRAKER v2.1.63, which trained the Augustus v3.4.03 and GeneMark-ES/ET/EP 4.68_lic predictors with mapped transcriptome data and protein homology information (OrthoDB v11) (Hoff et al. 2016; Stanke et al. 2004; Brůna et al. 2020). Transcriptome-based predictions relied on RNA-seq data aligned to the assembly using Hisat2 v2.2.03 to produce BAM alignment files (Kim et al. 2019). Assembled transcript sequences were generated using StringTie v2.1.4 based on these alignments (Kovaka et al. 2019). Gene predictions based on protein homology and intron conservation were performed using GeMoMa v1.8, leveraging protein sequences from six insect species (*Drosophila melanogaster*, *Tribolium castaneum*, *Cloeon dipterum*, *Daphnia magna*, *Rhopalosiphum maidis*, *Zootermopsis nevadensis*) to enhance sensitivity (Keilwagen et al. 2019). Finally, MAKER combined these results through the EVidenceModeler (EVM) pipeline v1.1.1 (Haas et al. 2008).

For functional annotation of genes, Diamond v2.0.8 was used to query the UniProtKB database with a sensitive mode. Protein domains were predicted using InterProScan v5.48-83.0, which screened proteins against public databases, including Pfam, SMART, Gene3D, Superfamily, and the Conserved Domain Database (CDD) (Buchfink et al. 2015). Additionally, InterProScan and the eggNOG v5.0 database were utilized to annotate genes with Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, KEGG orthologous groups (KO), Reactome pathways, clusters of orthologous groups (COG), and Expression coherence (EC) (Letunic et al. 2019).

Acknowledgments

This work was supported by the National Natural Science Foundation of China to R.L. (grant number 32200359) and the China Postdoctoral Science Foundation to R.L. (grant number 2024M751749).

Data Availability

The finalized genome assembly has been submitted to GenBank under the accession number JAHKSX000000000. Associated raw sequencing datasets, including second-generation, third-generation, and Hi-C data, are available in GenBank with the BioProject accession PRJNA735533 and the BioSample accession SAMN19588419. Annotation files are hosted on Figshare and can be accessed at https://doi.org/10.6084/m9.figshare.28120991.

Literature Cited

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 6:11.

Bauernfeind E, Soldán T. 2012. The Mayflies of Europe (Ephemeroptera). Ollerup, Denmark: Apollo Books. p. 781.

Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinformatics. 2:lqaa026.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12:59–60.

Bushnell B. 2014. BBMap Download. SourceForge.net. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 27 Apr 2020).

Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. In: Kollmar M, editor. Gene prediction: Methods and protocols. New York: Humana. p. 161–177.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016a. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3:99–101.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016b. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3:95–98.

El Yaagoubi S, Errochdi S, Edegbene AO, Kassout J, Harrak R, El Alami M. 2024. Distribution patterns of Ephemeroptera, Plecoptera, and Trichoptera (EPT) species in the northwestern Rif: environmental and climate change impacts. Hydrobiologia. Dec 10:1–9.

Farr A, Macadam CR, Lab NH, Darwin Tree of Life Consortium. 2023a. The genome sequence of the Northern Summer Mayfly, *Siphlonurus alternatus* (Say, 1824). Wellcome Open Res. 8.

Farr A, Macadam CR, Lab NH, Darwin Tree of Life Consortium. 2023b. The genome sequence of the Large Brook Dun, *Ecdyonurus torrentis* (Kimmins, 1942). Wellcome Open Res. 8.

Farr A, Skipp SJ, Macadam CR, Price BW. 2023c. The genome sequence of the pond olive, *Cloeon dipterum*. Wellcome Open Res. 8:540.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AFA. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 117:9451–9457.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 32:767–769.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491.

Hu J, Wang Z, Liang F, Liu SL, Ye K, Wang DP. 2024a. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. Genomics Proteomics Bioinformatics. 22:1.

Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, Ruan J. 2024b. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol. 25:107.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44:D81–D89.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47:D309–D314.

Jacobus LM, Macadam CR, Sartori M. 2019. Mayflies (Ephemeroptera) and their contributions to ecosystem services. Insects. 10:6.

Jacobus LM, Salles FF, Price BEN, Pereira-da-Conceicoa L, Dominguez E, Suter PJ, Molineri C, Tiunova TM, Sartori M. 2021. Mayfly taxonomy (Arthropoda: Hexapoda: Ephemeroptera) during the first two decades of the twenty-first century and the concentration of taxonomic publishing. Zootaxa. 4979:25–30.

Kamsoi O, Ventos-Alfonso A, Casares F, Almudi I, Belles X. 2021. Regulation of metamorphosis in neopteran insects is conserved in the paleopteran *Cloeon dipterum* (Ephemeroptera). Proc Natl Acad Sci USA. 118:e2105272118.

Keilwagen J, Hartung F, Grau J. 2019. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. In: Kollmar M, editor. Gene prediction: Methods and protocols. New York: Humana. p. 161–177.

Khudhur SM, Shekha YA. 2020. Morphological and molecular identification of three genera of the family Heptageniidae (Ephemeroptera) from Ava Sheen branch/Greater Zab tributary, North of Iraq. Iraqi J Sci. 61:952–960.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37:907–915.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20:278.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li R, Zhang WJ, Ma ZX, Zhou CF. 2020. Novel gene rearrangement pattern in the mitochondrial genomes of *Torleya mikhaili* and *Cincticostella fusca* (Ephemeroptera: Ephemerellidae). Int J Biol Macromol. 165:3106–3114.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of k-mers. Bioinformatics. 27:764–770.

Mayorga A, Lim C, Bae YJ. 2024. Use of mandibular tusks as weapons in the aggressive behavior of the burrowing mayfly *Rhoenanthus coreanus* (Yoon and Bae, 1985) (Ephemeroptera: Potamanthidae). Aquatic Insects. 45:49–59.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 29:2933–2935.

Rio DC, Ares M, Hannon GJ, Nilsen TW. 2010. Purification of RNA using TRIzol (TRI reagent). Cold Spring Harb Protoc. 2010:5439.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19:460.

Sartori M, Brittain JE. 2015. Order Ephemeroptera. In: Thorp JH, Rogers DC, editors. Ecology and General Biology, Vol I: Thorp and Covich's Freshwater Invertebrates. 4th ed. New York: Academic Press. p. 873–891.

Smit AFA, Hubley R, Green P. 2013–2015. Repeat Masker Open-4.0. Available from: http://www.repeatmasker.org.

Srinivasan P, Sivaruban T, Barathy S, Isack R. 2024. New findings of the *Caenis ulmeriana*-group (Ephemeroptera: Caenidae) in the Western Ghats, India. J Insect Biodiversity Systematics. Apr 5.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32:W309–W312.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 33:2202–2204.

Wakimura K, Ishiwata SI, Hamamura M, Takemon Y, Tanida K, Inai K, Kato M. 2024. Cox1-based phylogeny of Eastern Palearctic *Drunella* (Ephemeroptera: Ephemerellidae), description of new species and redescription of *D. cryptomeria* (Imanishi). Syst Biodivers. 22:2383214.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 35:543–548.

Zheng XHY, Gong DW, Qiang X, Zhou CF. 2024. A new genus and a new species of Behningiidae from China contributing new insights to the evolution of the family (Insecta: Ephemeroptera). Oriental Insects. Feb 7:1–32.

Table 1 Genome assembly and annotation statistics of *P. youi*.

Elements	**Value**
Genome assembly	
Assembly size (Mb)	412.902
Number of scaffolds/contigs	219/1429
Longest scaffold/contig (Mb)	45.886/2.902
N50 scaffold/contig length (Mb)	41.946/0.544
GC (%)	28.05
Gaps (%)	0.029
BUSCO completeness (%)	96.7
Gene annotation	
Protein-coding genes	15,647
Mean protein length (aa)	537.8
Mean gene length (bp)	9,934.7
Exons/introns per gene	11.2/9.9
Exon (%)	11.58
Mean exon length	270.9
Intron (%)	26.07
Mean intron length	694.6
BUSCO completeness (%)	97.6

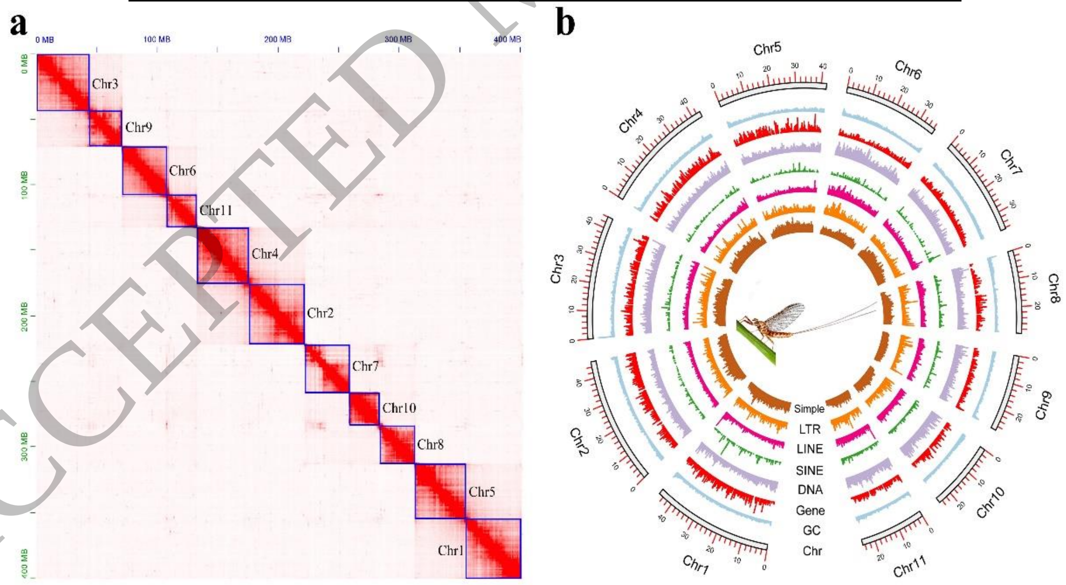

Figure 1

559x285 mm (x DPI)