scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Chromosome-level genome assembly of the stonefly *Indonemoura scalprata* (Plecoptera: Nemouridae)

Raorao Mo[1,2], Zhishan Cao[1,2], Jinjun Cao[1,2], Dávid Murányi[3], Ding Yang[4], Weihai Li[1,2] & Aili Lin[1,2] ✉

**Plecoptera (stoneflies) are hemimetabolous insects of significant ecological importance in freshwater ecosystems and are key to understanding early insect diversification and aquatic adaptation. Here, we present a high-quality, chromosome-level reference genome assembly for *Indonemoura scalprata* (Li & Yang, 2007), a nemourid stonefly widely distributed in southern China (Fujian, Guangdong, and Guangxi), with the sequenced specimen originating from the Qianjiadong National Nature Reserve in Guangxi. The genome was assembled using PacBio HiFi reads and Hi-C chromatin interaction data. The final assembly spans 257.81 Mb, with 99.76% (257.18 Mb) successfully anchored into 11 pseudochromosomes. The scaffold N50 reaches 26.45 Mb, and BUSCO completeness using the insecta_odb12 dataset (n = 3,114) is 97.0%. Repetitive elements constitute 31.20% of the genome (80.44 Mb). We identified 1,491 non-coding RNAs and predicted 12,200 protein-coding genes, achieving 97.7% completeness in the BUSCO (insecta_odb12) assessment. This high-quality genome provides a valuable resource for comparative genomics in Polyneoptera and studies of adaptation to cold, fast-flowing streams and early winged insect evolution.**

## Background & Summary

Plecoptera, commonly known as stoneflies, are a key component of EPT taxa (Ephemeroptera, Plecoptera, Trichoptera). Globally, approximately 17 families and 4,000 species are currently recognized, distributed across all continents except Antarctica[1,2], with their evolutionary origin dating back to the Upper Carboniferous period around 300 million years ago[3]. Stonefly nymphs typically inhabit clean, fast-flowing, well-oxygenated streams or springs, exhibiting an extensive altitudinal range from tens of meters above sea level to alpine snow zones exceeding 5,000 meters. Adults are generally terrestrial, feeding and mating in riparian habitats near the water bodies from which they emerged. Adult stoneflies typically do not feed or feed only sparingly, whereas nymphs occupy diverse trophic roles as predators, shredders, collector-gatherers, and scrapers, frequently exhibiting generalist feeding strategies or ontogenetic shifts in diet[4]. Mature nymphs usually emerge from the water onto stones or vegetation to undergo the final molt into adults, a process typically cued by temperature and/or photoperiod[5–9]. Due to their high sensitivity to water quality changes, Plecoptera are considered one of the most threatened freshwater insect groups by environmental change and are widely used as bioindicators in aquatic biomonitoring programs worldwide[10–14].

The family Nemouridae belongs to the order Plecoptera, suborder Arctoperlaria, infraorder Euholognatha, and superfamily Nemouroidea, and is the most species-rich and diverse family within the superfamily, currently comprising 21 genera and nearly 800 species—accounting for approximately half of the total species diversity in Nemouroidea[1]. As an ecologically significant group within the Plecoptera, its members are primarily found in lotic habitats and cold-water lakes within mountain stream ecosystems. This family occupies a key position

[1]Henan International Joint Laboratory of Taxonomy and Systematic Evolution of Insecta, Henan Institute of Science and Technology, Xinxiang, Henan, 453003, China. [2]Observation and Research Station on Water Ecosystem in Danjiangkou Reservoir of Henan Province, Nanyang, Henan, 474450, China. [3]Department of Zoology, Eszterházy Károly Catholic University, Leányka u. 6, Eger, H-3300, Hungary. [4]Department of Entomology, China Agricultural University, 2 Yuanmingyuan West Road, Beijing, 100193, China. ✉e-mail: linaili2023@126.com

in the phylogeny of Plecoptera and is divided into two subfamilies: Amphinemurinae and Nemourinae. The genus *Indonemoura* Baumann, 1975, belonging to the subfamily Amphinemurinae, is relatively species-rich, with over 60 species recorded worldwide, predominantly distributed in the Oriental Region, of which 35 species are known from China, representing a significant component of the genus diversity[15–17].

Phylogenetic analyses based on both morphological and molecular data consistently support the monophyly of Nemouridae[18–30]. Mitogenome-based phylogenetic analyses further indicate that Nemouridae and Notonemouridae form a sister group, a finding consistent with the morphology-based phylogenetic hypothesis for Plecoptera at the family level proposed by Zwick (2000)[21,24,25,27,28,30,31]. Moreover, the sister relationship between the two subfamilies is supported[24,31]. Currently, research on the biogeographic dispersal history, developmental and evolutionary trajectory, and genetic basis of adaptation in Nemouridae remains at a preliminary stage, primarily due to the scarcity of large-scale genomic data within this group. To resolve these phylogenetic conflicts and elucidate the evolutionary history of Nemouridae, high-quality genomic resources across its diverse lineages are urgently needed. However, such data remain scarce. To date, only 12 genomes of Plecoptera species have been deposited globally (accessed from the NCBI database in August 2025), among which four belong to the family Nemouridae. Within this family, two species from the subfamily Nemourinae have been assembled at the chromosome level and formally published[32,33]. Although no genomes from the subfamily Amphinemurinae have been officially published, one species has achieved chromosome-level assembly.

In this study, we obtained a high-quality chromosome-level genome assembly for *Indonemoura scalprata* (Li & Yang, 2007), a species representative of the data-deficient Amphinemurinae subfamily, distributed in Fujian, Guangdong, and Guangxi of China[15,16,34–36], using PacBio HiFi, short-read sequencing (for genomic survey and RNA-seq), Hi-C, and long-read transcriptome sequencing. We annotated essential genomic elements, including repetitive elements, non-coding RNAs (ncRNAs), and protein-coding genes. This genomic resource provides a valuable foundation for studying the biology, evolution, and conservation of stoneflies and their unique adaptations to freshwater ecosystems.

## Methods

**Sample collection.** Adult specimens of *Indonemoura scalprata* used in this study were collected on November 25, 2024, from the Qianjiadong National Nature Reserve in the Guangxi Zhuang Autonomous Region, China (25°25′28″ N, 111°12′15″ E, 404 m; and 25°28′45″ N, 111°12′26″ E, 290 m). The samples were flash-frozen in liquid nitrogen and stored at −80 °C until DNA extraction. Following DNA extraction, the voucher specimens were deposited in the Insect Collection of the Henan Institute of Science and Technology (HIST), Xinxiang, Henan Province, China.

**Genome and transcriptome sequencing.** PacBio HiFi, genome survey (WGS), and Hi-C sequencing were performed on individual specimens. For transcriptome analysis, both short-read RNA-seq and full-length ONT transcriptome sequencing were conducted using RNA extracted from a single specimen. Genomic DNA and total RNA for standard RNA-seq were extracted from the specimen using a CTAB-based method and TRIzol™ Reagent, respectively. For full-length transcriptome sequencing, RNA was isolated using the RNA prep Pure Plant Plus Kit (DP441).

A 20-kb insert SMRTbell® library was constructed using the SMRTbell® Express Template Prep Kit 3.0 (Pacific Biosciences) for PacBio HiFi genome sequencing. An ONT full-length transcriptome library was independently prepared using the SQK-PCS109 and SQK-PBK004 kits (Oxford Nanopore Technologies) for isoform-level analysis.

A 350-bp insert library was constructed using the VAHTS Universal Plus DNA Library Prep Kit (Vazyme) and sequenced on an Illumina Xplus platform with 150-bp paired-end reads.

In brief, Hi-C libraries were generated using the restriction enzyme MboI (GATC recognition site), followed by proximity ligation and amplification. Chromatin was crosslinked with formaldehyde and digested with MboI. After biotin labeling of the digested ends, proximity ligation was performed, and biotinylated ligation products were enriched, sheared, and converted into Illumina-compatible libraries through end repair, A-tailing, adapter ligation, and PCR amplification.

Berry Genomics was responsible for constructing the PacBio HiFi, Illumina WGS, Illumina RNA-seq, and Hi-C libraries. BenaGen constructed the ONT full-length transcriptome library.

Finally, we generated 23.46 Gb (91.01X) of PacBio HiFi reads, 12.12 Gb (47.01X) of Illumina WGS reads, and 35.46 Gb (137.53X) of Hi-C reads for *de novo* genome assembly (Table 1). For transcriptome analysis, two RNA libraries were sequenced: one short-read RNA-seq library on the Illumina Xplus platform yielding 9.63 Gb of data, and one full-length transcriptome library on the Oxford Nanopore PromethION platform generating 12.60 Gb of long-read RNA sequences. These multi-platform sequencing data provide comprehensive support for high-quality genome assembly, gene annotation, and functional genomics analysis of *Indonemoura scalprata*.

**Genome assembly and scaffolding.** The Illumina short-read data were subjected to quality control and trimming using Fastp version 0.23.4[37] with the following parameters: quality trimming (>Q20), removal of duplicate reads (-D), poly-G/X tail trimming (-g -x), exclusion of reads with more than 10% unqualified bases (-u 10), and base correction using overlapping read information (-c). We analyzed the genomic features of *Indonemoura scalprata* using a k-mer approach, with k-mer counts generated by BBTools v38.90[38] (k = 21). Genome characteristics were estimated using GenomeScope2 version 2.0[39] with a maximum k-mer coverage cutoff of 1,000. The analysis predicted a haploid genome size of 264.75 Mb, a heterozygosity rate of 2.51%, and a repeat content of 16.68% (Fig. 1).

For *de novo* genome assembly, high-fidelity PacBio HiFi reads were generated using pbccs v6.4.0 (https://github.com/PacificBiosciences/ccs) with a minimum quality score of Q20. Initial assembly was performed with

| Library Type | Insert Size | Sequencing Platform | Read Type | Clean Data (Gb) | Sequencing Coverage (X) | SRA Accession Number |
|---|---|---|---|---|---|---|
| WGS | 350 bp | Illumina Xplus | PE150 | 12.12 | 47.01 | SRR35196256 |
| HiFi | ~20 kb | PacBio Sequel Revio | HiFi long reads | 23.46 | 91.01 | SRR35196254 |
| Hi-C | 300–800 bp | Illumina Xplus | PE150 | 35.46 | 137.53 | SRR35196255 |
| RNA-sr | 350 bp | Illumina Xplus | PE150 | 9.63 | — | SRR35196252 |
| RNA-ONT | N/A (full-length cDNA) | Oxford Nanopore PromethION | Long reads | 12.60 | — | SRR35196253 |

**Table 1.** Sequencing data for *de novo* genome assembly and annotation of *Indonemoura scalprata*. Note: PE150 denotes paired-end 150 bp reads. *de novo* assembly was supported by PacBio HiFi and Hi-C data.

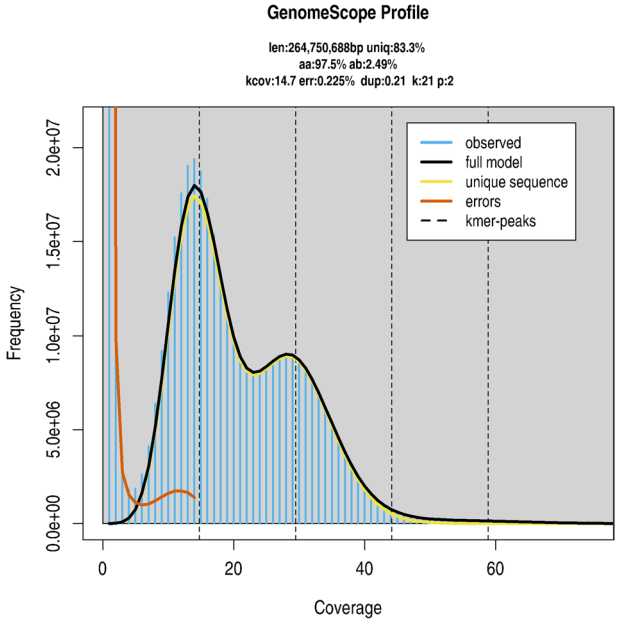


**Fig. 1** GenomeScope genome size estimates for *Indonemoura scalprata*.

Hifiasm v0.24.0[40] using default parameters, excluding contigs with coverage below 9X, corresponding to less than 1/10 of the average sequencing depth. Redundant contigs resulting from heterozygosity were removed using Purge_dups v1.2.5[41] with default parameters (−2 -a 70), based on contig similarity and sequencing depth. Briefly, HiFi reads were first aligned to the initial assembly using Minimap2 v2.29[42,43] with the -x map-hifi preset to generate a PAF file for Purge_dups input; Minimap2 was also used for self-alignment of the assembly (-x asm5 -DP) for downstream analyses. The assembly was not polished with short reads due to the inherently high accuracy of HiFi data, as confirmed by Merqury analysis showing a consensus quality value (QV) >60.

Chromosome-level scaffolding was achieved using Hi-C data. Briefly, Hi-C reads were aligned and processed using Chromap v0.2.6[44], with duplicate removal and Hi-C contact extraction, followed by initial scaffolding with YAHS v1.2[45]. Manual correction of misjoins, inversions, and translocations was performed using Juicebox v1.11.0810[46], after which a second round of scaffolding was conducted to generate the final assembly for *I. scalprata*. The sequencing depth of each pseudochromosome was independently assessed using SAMtools v1.10[47] based on alignments generated by Minimap2 with either PacBio HiFi reads (-ax map-hifi) or short-read WGS data (-ax sr). The final assembly yielded 11 pseudochromosomes ranging from 14.92 Mb to 36.69 Mb in length, with a total size of 257.18 Mb (99.76% of the assembled genome; Figs. 2, 3; Table S1). The circular visualization was generated using TBTools (v2.096). These chromosomes show high and uniform sequencing coverage and QV values exceeding 68, supporting a high-quality, chromosome-level genome assembly (Table S1).

Assembly quality was evaluated using multiple metrics. BUSCO v6.0.0[48] with the insecta_odb12 dataset (n = 3,114) showed 97.0% completeness, with only 0.8% duplicated genes, indicating high completeness and low redundancy. Read mapping rates, calculated from alignments generated by Minimap2 and SAMtools, were exceptionally high: 93.98% (Illumina WGS), 96.11% (RNA-seq), 96.34% (ONT full-length RNA), and 99.89% (PacBio HiFi), confirming high data utilization and assembly integrity. Contamination screening was conducted using MMseq. 2 v13[49] against the NCBI nt and UniVec databases, revealing no significant non-target sequences. Merqury v1.3[50] confirmed high base accuracy (QV >60) and low heterozygosity across all chromosomes.

The final chromosome-level assembly of *I. scalprata* spans 257.81 Mb, with 57 contigs and 22 scaffolds, a scaffold N50 of 26.45 Mb, and the largest scaffold reaching 36.69 Mb (Table 2). The GC content is 36.39% (Table 2). This assembly exhibits high contiguity and accuracy, representing a high-quality reference genome for the stonefly lineage. Compared to the four published and assembled Plecoptera genomes, the genome size of *I. scalprata* is

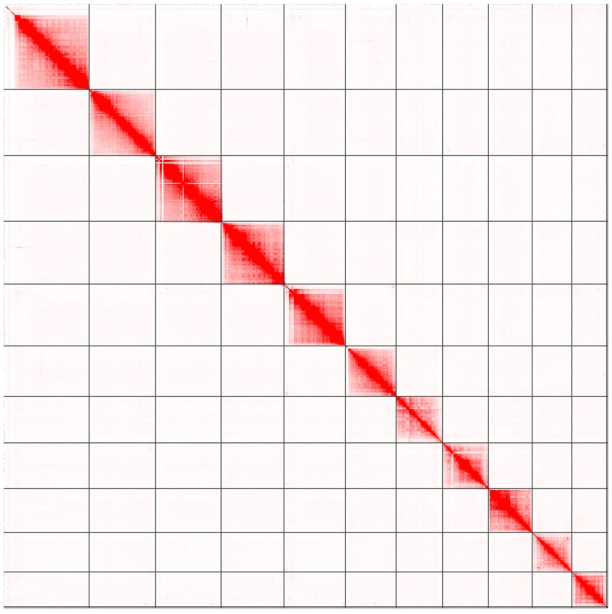

**Fig. 2** Genome-wide chromosomal heatmap of *Indonemoura scalprata*.

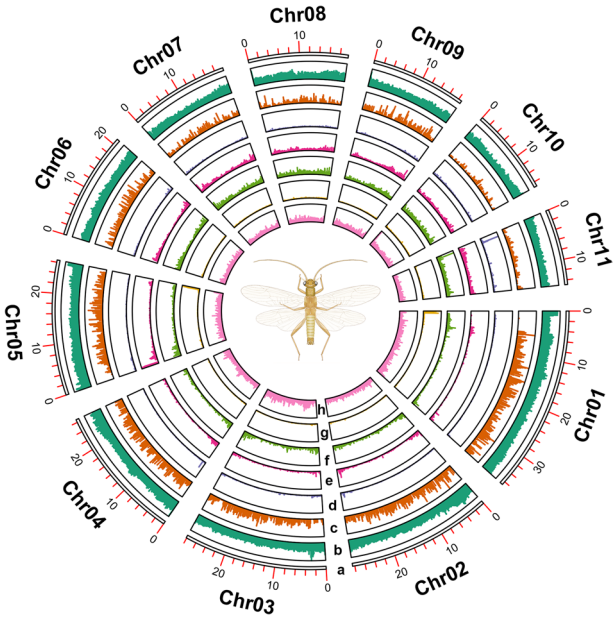


**Fig. 3** The genome characteristics of *Indonemoura scalprata* are depicted in a circular layout. Blocks on the outermost circle represent the 11 pseudochromosomes of *I. scalprata*. Peak plots from outer to inner circles represent chromosome length (**a**), GC content (**b**), protein-coding gene density (**c**), DNA transposon density (**d**), SINE density (**e**), LINE density (**f**), LTR density (**g**), and simple repeats density (**h**), respectively.

smaller than that of *Brachyptera putata* (Taeniopterygidae; 436.5 Mb)[51], *Leuctra nigra* (Leuctridae; 536.3 Mb)[52], and *Nemoura dubitans* (Nemouridae; 321.0 Mb)[33], but similar to *Nemurella picteti* (Nemouridae; 257 Mb)[32]. Its GC content (36.39%) is also lower than that of the other species (*B. putata*: 36.5%, *L. nigra*: 38%, *N. dubitans*: 43.5%, and *N. picteti*: 43.5%).

**Genome annotation.** Repetitive elements in the genome of *Indonemoura scalprata* were identified using a combined approach of homology-based and *de novo* prediction methods. A species-specific *de novo* repeat library was constructed using RepeatModeler v2.0.5[53] with the LTR structural search pipeline (-LTRStruct). This *de novo* library was merged with the Dfam 3.8[54] and RepBase-20181026[55] databases to generate a custom, comprehensive repeat database. Repetitive sequences were identified by homology-based alignment using RepeatMasker v4.1.5[56] and the custom repeat library. A total of 80.44 Mb of repetitive sequences were identified, accounting for

| Characteristics | *I. scalprata* |
|---|---|
| **Genome Assembly** | |
| Assembly size (Mb) | 257.81 |
| Number of scaffolds / contigs | 22 / 57 |
| Longest scaffold / contig (Mb) | 36.692 / 33.232 |
| Scaffold N50 / Contig N50 (Mb) | 26.446 / 16.69 |
| GC content (%) | 36.39 |
| Chromosome-level anchoring (Mb, %) | 257.18 (99.76%) |
| **BUSCO Completeness (%)** | |
| Complete | 97.0 |
| Complete single-copy | 96.3 |
| Complete duplicated | 0.8 |
| Fragmented | 0.3 |
| Missing | 2.7 |
| **Read Mapping Rate (%)** | |
| Illumina WGS | 93.98 |
| PacBio HiFi | 99.89 |
| RNA-seq (short-read) | 96.11 |
| RNA-ONT (long-read) | 96.34 |

**Table 2.** Genome assembly summary for *Indonemoura scalprata*.

31.20% of the *I. scalprata* genome. This includes 29.72% interspersed repeats, 1.12% simple repeats, 0.26% satellites, 0.09% low-complexity regions, and 0.01% rRNA/tRNA/snRNA-derived repeats. Among the interspersed repeats, 3.07% of the genome were classified as LINEs, 2.43% as DNA transposons, 3.24% as LTR retrotransposons, 1.23% as Rolling-circle elements, and 1.74% as SINEs. Notably, unclassified repeats constituted 17.98% of the genome, indicating a substantial proportion of novel or divergent repetitive elements (Table S2). Gypsy (0.80%), L2 (1.44%), and Helitron (1.23%) were the most prevalent known TE families. Consistent with a history of recent transposable element activity, the distribution of Kimura 2-parameter divergence values showed a peak at low divergence levels, suggesting a potential burst of TE expansion in the *I. scalprata* genome (Fig. S1).

Two strategies were employed for noncoding RNA annotation in the *I. scalprata* genome: (1) homology-based identification using the Rfam database and Infernal version 1.1.5[57] for structural RNAs, and (2) *de novo* prediction of transfer RNAs using tRNAscan-SE version 2.0.12[58] with filtering of low-confidence hits via the built-in EukHighConfidenceFilter script. Noncoding RNAs, including transfer RNAs (tRNAs), microRNAs (miRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs), were identified accordingly. A total of 1,491 noncoding RNAs (ncRNAs) were annotated, encompassing both infrastructural (housekeeping) and regulatory RNA classes. The numbers of rRNA, tRNA, snRNA, miRNA, lncRNA, ribozyme, and other ncRNA motifs were 605, 402, 87, 54, 2, 3, and 341, respectively. The rRNAs included four 5.8S rRNAs, 583 5S rRNAs, eight large subunit (LSU) rRNAs, and ten small subunit (SSU) rRNAs. The tRNAs covered 21 amino acid isotypes, with the highest copy number observed for tRNA-Pro (36 copies) and the lowest for tRNA-SeC (1 copy). The snRNAs were classified into 55 spliceosomal RNAs (including U1, U2, U4, U5, U6, and U11), three minor spliceosomal RNAs (U4atac, U6atac, U12), 25 C/D box snoRNAs, and three H/ACA box snoRNAs. A total of 54 miRNA genes were identified, representing 44 distinct miRNA families based on Rfam annotation, including evolutionarily conserved ones such as *let-7*, *mir-1*, *mir-8*, and *mir-9*. Additionally, two long non-coding RNAs (Sphinx_1 and Sphinx_2) and three catalytic RNAs (two nuclear RNase P and one RNase MRP) were detected (Table S3). Notably, 341 copies of the histone 3′ UTR stem-loop motif were identified, reflecting the tandem organization of core histone genes in the genome.

Protein-coding genes were annotated by integrating evidence from (1) *ab initio* prediction, (2) transcriptome-based prediction, and (3) homology-based prediction. For *ab initio* prediction, BRAKER v3.0.6[59] and GeMoMa v1.9[60] were used to generate gene models, which were combined and supplied as the *ab initio* evidence for MAKER. BRAKER automatically trained Augustus v3.4.0[61] and GeneMark-ETP[62] using RNA-seq data and protein sequences from the OrthoDB11 database[63]. Short-read RNA-seq alignments were generated with minimap2 using the splice-aware option -x splice:sr. For transcriptome-based prediction, both short-read (Illumina) and long-read (Nanopore) RNA-seq data were assembled using StringTie v2.2.1[64] in–mix mode, with splice junctions guided by the genome alignment BAM files produced by minimap2 (-x splice). For homology-based prediction, GeMoMa was employed to transfer annotations from high-quality genomes of closely related species, including *Ischnura elegans* (Odonata), *Drosophila melanogaster* (Diptera), *Rhopalosiphum maidis* (Hemiptera), and three polyneopteran insects: *Anabrus simplex* (Orthoptera), *Bacillus rossius redtenbacheri* (Phasmatodea), and *Periplaneta americana* (Blattaria). GeMoMa was run with parameters GeMoMa.c = 0.4 and GeMoMa.m = 57,000 for *I. scalprata*. All three lines of evidence were integrated using the MAKER v3.01.04[65] pipeline to produce the final gene models. A total of 12,200 protein-coding genes were predicted, with a mean gene length of 11,407.7 bp and a mean protein length of 588.5 amino acids. On average, each gene contained 8.2 exons (mean length: 299.8 bp) and 7.2 introns (mean length: 1,309.0 bp), corresponding to 7.9 CDSs per gene (mean length: 218.2 bp) (Table 3). The gene space completeness was assessed using BUSCO

| Annotation | Number |
|---|---|
| **Structure annotation** | |
| Number of protein-coding genes | 12,200 |
| Number of predicted protein sequences | 15,638 |
| Mean protein length (aa) | 588.5 |
| Mean gene length (bp) | 11,407.70 |
| Gene ratio | 53.98% |
| Number of exons per gene | 8.2 |
| Mean exon length (bp) | 299.8 |
| Exon ratio | 11.72% |
| Number of CDSs per gene | 7.9 |
| Mean CDS length (bp) | 218.2 |
| CDS ratio | 8.23% |
| Number of introns per gene | 7.2 |
| Mean intron length (bp) | 1,309.00 |
| Intron ratio | 42.26% |
| **Function annotation** | |
| Number of genes matching Uniprot records | 11,343 |
| Number of genes labelled as "Uncharacterized protein" | 588 |
| Number of genes labelled as "unknown function" | 903 |
| Number of genes with InterProScan annotations | 10,318 |
| Number of genes with GO items from InterProScan annotations | 6,273 |
| Number of genes with eggNOG annotations | 11,306 |
| Number of genes with GO items from eggNOG annotations | 8,599 |
| Number of genes with Enzyme Codes (EC) from eggNOG annotations | 2,757 |
| Number of genes with KEGG ko terms from eggNOG annotations | 7,743 |
| Number of genes with KEGG pathway terms from eggNOG annotations | 4,744 |
| Number of genes with COG Functional Categories from eggNOG annotations | 10,699 |
| Number of genes with GO items (combining InterProScan and eggNOG results) | 9,679 |
| Number of genes with KEGG pathways items (combining InterProScan and eggNOG results) | 4,744 |

**Table 3.** Summary of protein-coding gene structural features in the *Indonemoura scalprata* genome.

(insecta_odb12, n = 3,114), yielding a completeness score of C:97.7% [S:81.6%, D:16.1%], F:0.7%, M:1.6%, indicating high-quality gene annotation.

Gene functional annotation was performed by searching against the UniProtKB (Swiss-Prot + TrEMBL) database (https://www.uniprot.org/) using DIAMOND v2.1.7.161[66] with the '–very-sensitive' mode and an E-value cutoff of 1e-5. Protein domains and Gene Ontology (GO) terms were assigned using InterProScan v5.70-102.0[67] against multiple databases, including Pfam (v32.0)[68], SMART, Gene3D, SuperFamily, and the Conserved Domain Database (CDD). Additionally, eggNOG-mapper v2.1.12[69] was employed with the eggNOG v5.0.2[70] database to annotate orthologous groups, GO terms, KEGG orthology (KO), COG functional categories, and Enzyme Commission (EC) numbers. The results from InterProScan and eggNOG-mapper were integrated to generate comprehensive functional annotations for the protein-coding genes. A total of 11,343 (92.98%) protein-coding genes showed significant homology to entries in the UniProtKB database. Functional domain analysis revealed that 10,318 (84.59%) genes were assigned at least one domain by InterProScan. When combining InterProScan and eggNOG-mapper results, 9,679 (79.33%) genes were annotated with Gene Ontology (GO) terms, and 4,744 (38.89%) were assigned KEGG pathway entries (Table 3). The overall functional annotation coverage in the *I. scalprata* genome is comparable to other well-annotated insect genomes. The high annotation rate, particularly for protein domains (84.57%), indicates a robust and sensitive annotation pipeline.

## Data Records

The raw sequencing data and genome assembly of *Indonemoura scalprata* have been deposited under BioProject PRJNA1234515 at the National Center for Biotechnology Information (NCBI). Accession numbers for the data are as follows: PacBio HiFi (SRR35196254)[71], Illumina WGS (SRR35196256)[72], Hi-C (SRR35196255)[73], Illumina RNA-seq (SRR35196252)[74], and Oxford Nanopore full-length transcriptome (SRR35196253)[75]. The assembled genome has been deposited in the NCBI GenBank database under accession number JBMUJJ000000000[76]. Additionally, the genome annotation results have been deposited in the Figshare repository[77].

## Technical Validation

The quality of the *Indonemoura scalprata* genome assembly was assessed using multiple complementary approaches. First, completeness was evaluated using BUSCO v6.0.0 with the insecta_odb12 dataset (n = 3,114), yielding a high completeness score of 97.0% (S: 96.3%, D: 0.8%) at the genome level, indicating excellent gene space representation and low redundancy. Second, assembly accuracy was assessed by mapping sequencing

reads back to the assembly. Mapping rates were exceptionally high: 99.89% for PacBio HiFi reads, 93.98% for Illumina WGS reads, 96.11% for Illumina RNA-seq reads, and 96.34% for Oxford Nanopore full-length transcriptome reads, demonstrating strong sequence-level consistency. Additionally, contamination screening using MMseq. 2 against the NCBI nt and UniVec databases revealed no significant non-target sequences. These results collectively confirm that the *I. scalprata* genome assembly is highly complete, accurate, and free of contamination, representing a high-quality reference for stonefly genomics.

## Data availability

The dataset described in this study is novel and has not been published previously. All data are publicly available. The raw sequencing data and genome assembly for *Indonemoura scalprata* have been deposited in the National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA1234515. Comprehensive genome annotation files, including repetitive element annotations, gene models, and functional annotations, are available in the Figshare repository at https://doi.org/10.6084/m9.figshare.30048700.v2.

## Code availability

No custom scripts were developed for this study. All data processing was performed using standard bioinformatics software, following the recommended protocols and official documentation, with parameter settings detailed in the Methods section.

Received: 10 September 2025; Accepted: 9 December 2025;
Published online: 17 December 2025

## References

1. DeWalt, R. E., Hopkins, H., Neu-Becker, U. & Stueber, G. Plecoptera Species File. https://plecoptera.speciesfile.org/ (accessed on 1 September) (2025).
2. Eichert, A. *et al*. Stonefly systematics: past, present, and future. *Insect Syst. Diver.* **9**, ixaf026, https://doi.org/10.1093/isd/ixaf026 (2025).
3. Béthoux, O., Cui, Y. Y., Kondratieff, B. C., Stark, B. P. & Ren, D. At last, a Pennsylvanian stem-stonefly (Plecoptera) discovered. *BMC Evol. Biol.* **11**, 248, https://doi.org/10.1186/1471-2148-11-248 (2011).
4. Tierno de Figueroa, J. M. & López-Rodriguez, M. J. Tropical ecology of Plecoptera (Insecta): a review. *Eur. Zool. J.* **86**, 79–102, https://doi.org/10.1080/24750263.2019.1592251 (2019).
5. Brittain, J. E. Life History Strategies in Ephemeroptera and Plecoptera, Dordrecht, 1–12.
6. Hynes, H. B. N. Biology of Plecoptera. *Annu. Rev. Entomol.* **21**, 135–153 (1976).
7. Stewart, K. W. & Stark, B. P. *Nymphs of North American Stonefly Genera (Plecoptera). 2nd Edition* (The Caddis Press, 2002).
8. DeWalt, R. E., Kondratieff, B. C. & Sandberg, J. B. Order Plecoptera, in *Thorp and Covich's Freshwater Invertebrates: ecology and general biology*, (ed. Thorp, J. & Rogers, D. C.) (Academic Press, 2015).
9. Wolf, B. & Zwick, P. Plurimodal emergence and plurivoltinism of Central European populations of *Nemurella pictetii* (Plecoptera: Nemouridae). *Oecologia.* **79**, 431–438, https://doi.org/10.1007/BF00378657 (1989).
10. Marwein, I. & Gupta, S. Plecoptera community of two small streams of Shillong, Meghalaya, North-East India. *Asian J. Conserv. Biol.* **10**, 28–39, https://doi.org/10.53562/ajcb.ENDY4688 (2021).
11. Kopec, A. D. *et al*. Spatial and temporal trends of mercury in the aquatic food web of the lower Penobscot River, Maine, USA, affected by a chlor-alkali plant. *Sci. Total Environ.* **649**, 770–791, https://doi.org/10.1016/j.scitotenv.2018.08.203 (2019).
12. DeWalt, R. E. & Ower, G. D. Ecosystem services, global diversity, and rate of stonefly species descriptions (Insecta: Plecoptera). *Insects*. **10**, 99, https://doi.org/10.3390/insects10040099 (2019).
13. DeWalt, R. E., Favret, C. & Webb, D. W. Just How Imperiled Are Aquatic Insects? A Case Study of Stoneflies (Plecoptera) in Illinois. *Ann. Entomol. Soc. Am.* **98**, 941–950, https://doi.org/10.1603/0013-8746(2005)098[0941:Jhiaai]2.0.Co;2 (2005).
14. Sánchez-Bayo, F. & Wyckhuys, K. A. G. Worldwide decline of the entomofauna: A review of its drivers. *Biol. Conserv.* **232**, 8–27, https://doi.org/10.1016/j.biocon.2019.01.020 (2019).
15. Yang, D., Li, W. H. & Zhu, F. *Fauna Sinica, Insecta Vol. 58, Plecoptera: Nemouroidea* (Science Press, 2015).
16. Yang, D. & Li, W. H. *Plecoptera. Species Catalogue of China. Vol. 2. Animals, Insecta (III)* (Science Press, 2018).
17. Baumann, R. W. Revision of the stonefly family Nemouridae (Plecoptera): A study of the world fauna at the generic level. *Smithson. Contrib. Zool.* **211**, 1–74, https://doi.org/10.5479/si.00810282.211 (1975).
18. Illies, J. Katalog der rezenten Plecoptera. *Das Tierreich.* **82**, 1–632 (1966).
19. Zwick, P. Insecta: Plecoptera Phylogenetisches System und Katalog. *Das Tierreich.* **94**, 1–465 (1973).
20. Zwick, P. Plecoptera (Steinfliegen). *Handb. Zool.* **4**, 1–115 (1980).
21. Zwick, P. Phylogenetic system and zoogeography of the Plecoptera. *Annu. Rev. Entomol.* **45**, 709–746 (2000).
22. Chen, Z. T. & Du, Y. Z. Complete mitochondrial genome of *Capnia zijinshana* (Plecoptera: Capniidae) and phylogenetic analysis among stoneflies. *J. Asia-Pac. Entomol.* **20**, 305–312, https://doi.org/10.1016/j.aspen.2017.01.013 (2017).
23. Chen, Z. T. & Du, Y. Z. First mitochondrial genome from Nemouridae (Plecoptera) reveals novel features of the elongated control region and phylogenetic implications. *Int. J. Mol.Sci.* **18**, 996, https://doi.org/10.3390/ijms18050996 (2017).
24. Cao, J. J., Wang, Y. & Li, W. H. Comparative mitogenomic analysis of species in the subfamily Amphinemurinae (Plecoptera: Nemouridae) reveal conserved mitochondrial genome organization. *Int. J. Biol. Macromol.* **138**, 292–301, https://doi.org/10.1016/j.ijbiomac.2019.07.087 (2019).
25. Cao, J. J. *et al*. Two Complete Mitochondrial Genomes From Leuctridae (Plecoptera: Nemouroidea): Implications for the Phylogenetic Relationships Among Stoneflies. *J. Insect Sci.* **21**, 1–6, https://doi.org/10.1093/jisesa/ieab009 (2021).
26. Chen, S. J., Cao, J. J., Li, W. H. & Wang, Y. Characterization of the complete mitochondrial genome of *Nemoura meniscata* Li & Yang (Plecoptera: Nemouridae) from China. *Mitochondrial DNA B.* **5**, 1052–1053, https://doi.org/10.1080/23802359.2020.1721370 (2020).
27. Ding, S. M. *et al*. The phylogeny and evolutionary timescale of stoneflies (Insecta: Plecoptera) inferred from mitochondrial genomes. *Mol. Phylogenet. Evol.* **135**, 123–135, https://doi.org/10.1016/j.ympev.2019.03.005 (2019).
28. Guo, X. *et al*. Mitochondrial Genome of *Strophopteryx fasciata* (Plecoptera: Taeniopterygidae), with a Phylogenetic Analysis of Nemouroidea. *Genes.* **13**, 1116, https://doi.org/10.3390/genes13071116 (2022).
29. South, E. J. *et al*. Phylogenomics of the North American Plecoptera. *Syst. Entomol.* **46**, 287–305, https://doi.org/10.1111/syen.12462 (2021).
30. Cao, J. J., Wang, Y., Muranyi, D., Cui, J. X. & Li, W. H. Mitochondrial genomes provide insights into the Euholognatha (Insecta: Plecoptera). *BMC Ecol. Evol.* **24**, 16, https://doi.org/10.1186/s12862-024-02205-6 (2024).

31. Wang, Y. *et al*. Mitochondrial genomes of Nemourinae species (Plecoptera: Nemouridae) and the phylogenetic implications. *J. Insect Sci.* **24**, 1–9, https://doi.org/10.1093/jisesa/ieae028 (2024).
32. Macadam, C., Clifford, C. & Price, B. W. The genome sequence of a stonefly, *Nemurella pictetii* Klapalek, 1900. *Wellcome Open Res.* **7**, https://doi.org/10.12688/wellcomeopenres.17684.1 (2022).
33. Farr, A. & Macadam, C. R. The genome sequence of a stonefly, *Nemoura dubitans* (Morton, 1894). *Wellcome Open Res.* **8**, https://doi.org/10.12688/wellcomeopenres.19771.1 (2023).
34. Li, W. H. & Yang, D. Review of the genus *Amphinemura* (Plecoptera: Nemouridae) from Guangdong, China. *Zootaxa.* **1511**, 55–64, https://doi.org/10.11646/zootaxa.1511.1.4 (2007).
35. Mo, R. R., Wang, G. Q., Yang, D., Li, W. H. & Murányi, D. Two new species and one new regional record of *Indonemoura* from Guangxi, China, with additions to larval characters (Plecoptera, Nemouridae). *Zookeys.* **825**, 25–42, https://doi.org/10.3897/zookeys.825.31663 (2019).
36. Mo, R. R., Wang, G. Q., Yang, D., Li, W. H. & Murányi, D. Two new species and four unknown larvae of Amphinemurinae (Plecoptera, Nemouridae) from southern China. *Zootaxa.* **5040**, 77–101, https://doi.org/10.11646/zootaxa.5040.1.4 (2021).
37. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
38. Bushnell, B. BBMap. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 October 2022) (2014).
39. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
40. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods.* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
41. Guan, D. F. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
42. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
43. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics.* **37**, 4572–4574, https://doi.org/10.1093/bioinformatics/btab705 (2021).
44. Zhang, H. W. *et al*. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat. Commun.* **12**, 6566, https://doi.org/10.1038/s41467-021-26865-w (2021).
45. Zhou, C. X., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics.* **39**, https://doi.org/10.1093/bioinformatics/btac808 (2022).
46. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
47. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience.* **10**, 1–4, https://doi.org/10.1093/gigascience/giab008 (2021).
48. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
49. Steinegger, M. & Söding, J. MMseqs. 2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028, https://doi.org/10.1038/nbt.3988 (2017).
50. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
51. Sawada, Y. *et al*. Unsupervised AI reveals insect species-specific genome signatures. *PeerJ.* **12**, e17025, https://doi.org/10.7717/peerj.17025 (2024).
52. Clifford, C., Macadam, C. R. & Price, B. W. The genome sequence of the black needle fly, *Leuctra nigra* (Olivier, 1811). *Wellcome Open Res.* **8**, https://doi.org/10.12688/wellcomeopenres.19082.1 (2023).
53. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci.* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
54. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA.* **12**, 2, https://doi.org/10.1186/s13100-020-00230-y (2021).
55. Bao, W. D., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA.* **6**, 11, https://doi.org/10.1186/s13100-015-0041-9 (2015).
56. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org (2013/2015).
57. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
58. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences, in *Gene Prediction: Methods and Protocols*, (ed. Kollmar, M.) (Springer New York, 2019).
59. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, https://doi.org/10.1093/nargab/lqaa108 (2021).
60. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics.* **19**, 189, https://doi.org/10.1186/s12859-018-2203-5 (2018).
61. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
62. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data. 2023.2001.2013.524024, https://doi.org/10.1101/2023.01.13.524024 (2024).
63. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451, https://doi.org/10.1093/nar/gkac998 (2023).
64. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
65. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* **12**, 491, https://doi.org/10.1186/1471-2105-12-491 (2011).
66. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods.* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
67. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427, https://doi.org/10.1093/nar/gkac993 (2022).
68. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432, https://doi.org/10.1093/nar/gky995 (2018).
69. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
70. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2018).
71. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35196254 (2025).

72. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35196256 (2025).
73. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35196255 (2025).
74. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35196252 (2025).
75. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35196253 (2025).
76. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_052425775.1 (2025).
77. Mo *et al*. The chromosome-level genome assembly of *Indonemoura scalprata* (Plecoptera: Nemouridae). *Figshare* https://doi.org/10.6084/m9.figshare.30048700.v2 (2025).

## Acknowledgements

We sincerely thank Prof. Feng Zhang for his insightful and constructive feedback on genome annotation and phylogenetic reconstruction. This work was supported by the National Natural Science Foundation of China (32400370 and 32270492) and the International Science and Technology Cooperation Project of Henan Province (242102521015).

## Author contributions

R.R.M., D.Y., W.H.L. and A.L.L. contributed to the research design. R.R.M., W.H.L. and A.L.L. collected the samples. R.R.M., Z.S.C., J.J.C. and A.L.L. analyzed the data. R.R.M., D.M. and A.L.L. wrote the draft manuscript. All authors contributed to revising the manuscript. All authors have been read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06430-6.

**Correspondence** and requests for materials should be addressed to A.L.