First Annotated Genome of a Mandibulate Moth, *Neomicropteryx cornuta*, Generated Using PacBio HiFi Sequencing

Xuankun Li1, Emily Ellis1, David Plotkin1, Yume Imada2, Masaya Yago3, Jacqueline Heckenhauer4,5, Timothy P. Cleland6, Rebecca B. Dikow7, Torsten Dikow8, Caroline G. Storer1, Akito Y. Kawahara1,*, and Paul B. Frandsen4,7,9,*

1McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, USA
2Graduate School of Science and Engineering, Ehime University, Matsuyama, Japan
3The University Museum, The University of Tokyo, Hongo, Bunkyo-ku, Japan
4LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany
5Department of Terrestrial Zoology, Entomology III, Senckenberg Research Institute and Natural History Museum Frankfurt, Frankfurt, Germany
6Museum Conservation Institute, Smithsonian Institution, Suitland, Maryland, USA
7Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, Washington, District of Columbia, USA
8Department of Entomology, National Museum of Natural History (USNM), Smithsonian Institution, Washington, District of Columbia, USA
9Department of Plant and Wildlife Sciences, Brigham Young University, USA

*Corresponding authors: E-mails: kawahara@flmnh.ufl.edu; paul_frandsen@byu.edu.

Accepted: 27 September 2021

Abstract

We provide a new, annotated genome assembly of *Neomicropteryx cornuta*, a species of the so-called mandibulate archaic moths (Lepidoptera: Micropterigidae). These moths belong to a lineage that is thought to have split from all other Lepidoptera more than 300 Ma and are consequently vital to understanding the early evolution of superorder Amphiesmenoptera, which contains the order Lepidoptera (butterflies and moths) and its sister order Trichoptera (caddisflies). Using PacBio HiFi sequencing reads, we assembled a highly contiguous genome with a contig N50 of nearly 17 Mb. The assembled genome length of 541,115,538 bp is about half the length of the largest published Amphiesmenoptera genome (*Limnephilus lunatus*, Trichoptera) and double the length of the smallest (*Papilio polytes*, Lepidoptera). We find high recovery of universal single copy orthologs with 98.1% of BUSCO genes present and provide a genome annotation of 15,643 genes aided by resolved isoforms from PacBio IsoSeq data. This high-quality genome assembly provides an important resource for studying ecological and evolutionary transitions in the early evolution of Amphiesmenoptera.

Key words: Amphiesmenoptera, HiFi, IsoSeq, Micropterigidae, Lepidoptera, PacBio.

Significance

*Neomicropteryx cornuta* is a member of the family Micropterigidae, sister to all other extant Lepidoptera. In this article, we report the first high-quality genome of a micropterigid, which is essential for studying ecological and evolutionary transitions in the early evolution of superorder Amphiesmenoptera.
Introduction

Lepidoptera are one of the most diverse herbivorous insect lineages, with more than 160,000 described species (Mitter et al. 2017). They are one of the two extant orders, along with caddisflies (Trichoptera), that comprise the superorder Amphiesmenoptera. Modern Lepidoptera and Trichoptera are morphologically similar in many ways, including dense covering of hairs or scales on their wings (Kristensen 1984), but in the ~310 Myr since they were believed to have diverged from each other (Kawahara et al. 2019), they have developed very different behaviors and ecological roles. Trichoptera larvae are primarily aquatic, with a diversity of feeding behaviors ranging from pure herbivory to opportunistic scavenging to predation (Mackay and Wiggins 1979). In contrast, almost all Lepidoptera larvae are terrestrial and herbivorous, and the adults provide essential pollination services for many flowering plants (Scoble 1992).

The earliest diverging lineage of Lepidoptera includes the mandibulate archaic moths, Micropterigidae (Kawahara et al. 2019). The family includes roughly 20 extant genera (Van Nieukerken et al. 2011), and is sister to all other extant Lepidoptera. Its fossil record dates back to the Lower Cretaceous (Azar et al. 2010; Kristensen and Skalski 1999; Whalley 1978), though recent phylogenetic studies estimate that the family could be as old as 300 Myr (Kawahara et al. 2019). Micropterigids are known for their unusual feeding habits and mouthpart morphology relative to other moths. The larvae feed on liverworts (Imada et al. 2011), whereas the larvae of most other extant Lepidoptera feed on angiosperms. Many micropterigid larvae, including those in the genus Neomicropteryx, have a plastron and other morphological features conducive to survival in flooded habitats (Davis and Landry 2012); this aquatic association is in sharp contrast with the primarily terrestrial habitats of most other Lepidoptera. Micropterigid adults have mandibulate (chewing) mouthparts, with some species feeding on angiospermous pollen (Kristensen 1999) or spores of ferns and lycopods (Gibbs 2014), whereas most other adult Lepidoptera are either non-feeding or consume nectar with an elongate, flexible proboscis (siphoning-sucking mouthparts). Fossil evidence suggests that the most recent common ancestor of Lepidoptera was mandibulate, like extant Micropterigidae, and had small structures called galea (also present in micropterigid) that evolved into the proboscis found in nearly all other butterflies and moths (Kristensen 1984; Krenn 2010). Since micropterigid remained mandibulate, their genetic makeup could shed light on the early evolution of ancient Lepidoptera and Trichoptera.

Despite their unique ecology and the fact that Micropterigidae represent a possible important transition between Trichoptera and Lepidoptera, there are no existing genome assemblies of Micropterigidae. By November 2020, there were 118 Lepidoptera and six Trichoptera genome assemblies available on GenBank (Hotaling et al. 2021). With more than 250 Myr of evolution between available genomes of Trichoptera and Lepidoptera (Triant et al. 2018), a Micropterigidae genome is an important evolutionary resource. Moreover, both orders are known for producing silk, but the structure and function of that silk can vary greatly between the two orders. Modern genomic analysis is an essential tool for extrapolating the evolutionary processes and transitions that resulted in the extant diversity stemming from the ancestral amphiesmenopteran. Here, we provide an annotated genome of Neomicropteryx cornuta, the first available genome of any mandibulate Lepidoptera.

We use PacBio HiFi sequencing data to assemble a highly contiguous N. cornuta genome. This is especially important since many genomes of Lepidoptera are of low quality (Ellis et al. 2021). We also provide a genome annotation by resolved isoforms from PacBio IsoSeq data. Our genome assembly provides an important resource to study ecological and evolutionary transitions in the early evolution of Amphiesmenoptera and sets the stage for future studies on the genomics of Amphiesmenoptera.

Results and Discussion

Assembly

Sequencing the N. cornuta genome using two PacBio SMRT cells produced 8.8 and 8.4 Gb of HiFi data, respectively, corresponding to ~31x PacBio HiFi read coverage. Blobtools analysis assigned 99.8% of all base pairs to the phylum Arthropoda (supplementary fig. 1, Supplementary Material online) and the resulting assembly contained 101 contigs with a contig N50 of 16,921,359 bp. This is the second-longest contig N50 for an amphiesmenopteran genome published thus far (Hotaling et al. 2021), shorter only than the genome of the silk moth Samia ricini (GCA_014132275.1), which was also generated by PacBio HiFi sequencing. Assembly GC content was 33.4% and the total assembly length was 541,115,538 bp, which is intermediate in length compared with other Amphiesmenoptera genomes (with BUSCO scores > 90%), which range from 227,005,758 bp (Papilio polytes) to 1,369,180,260 bp (Linneaphilus lunatus) (Hotaling et al. 2021). BUSCO analysis identified 98.1% (97.9% complete; 0.3% fragmented) of the Insecta gene set in the assembly (fig. 1, table 1).

Annotation

We also report the functional annotations of N. cornuta. Of the 15,643 predicted proteins, 86.62% (13,550) were verified by BLAST and/or transcript evidence, 63.04% (9,862) were mapped to GO terms, and 43.95% (6,875) were functionally annotated in Blast2Go. Top GO annotations include catalytic activity (4,512), cellular process (4,492), binding (4,414), and metabolic process (4,296) (supplementary figs. 4–6, Supplementary Material online). We annotated a total of
48.61% of the genome assembly as repeats using RepeatModeler and RepeatMasker. Unclassified repeats comprise >161 million bases, which is the highest among all types of repeats (genome proportion of 29.82%). Long interspersed nuclear elements (LINEs) are the second most abundant repeat category with >39 million bases (7.23%), followed by DNA transposons and rolling circles and long terminal repeats (LTRs), which have >15 million (2.82%) and >11 million bases (2.10%) respectively. Percent composition of repeats and predominance of LINEs were similar to both the S. ricini and Bombyx mori genome assemblies (Lee et al. 2021).

Conclusions

Our results provide a new genome for a relict evolutionary lineage, separated by more than 250 Myr of evolution from any currently existing genome. Results from our study show that high fidelity, long-read sequencing facilitates the production of more contiguous assemblies and generates high-quality resources for further investigation of genome functions. Our new genome will be useful for future studies on amphiesmenopteran genetics, conservation and ecology.

Materials and Methods

Sequencing and Assembly

Larval specimens of N. cornuta were field collected at two sites in Kochi Prefecture, Japan, and flash frozen (supplementary note 1, Supplementary Material online). DNA was extracted from a single specimen using a Zymo Quick-prep HMW DNA extraction kit. Following DNA extraction, the sequencing library was prepared according to the "Using Express Template Prep Kit 2.0 With Low DNA Input" protocol from PacBio. The library was then sequenced on two PacBio Sequel II SMRT cells in CCS mode. Further details are provided in supplementary note 1, Supplementary Material online. Q20 HiFi CCS reads were generated from the raw data using the pbccs tool, which is included in the pbbioconda package (https://github.com/PacificBiosciences/pbbioconda, last accessed August 9, 2021). The reads were then assembled into contigs using Hifiasm v0.13-r307 with the option for aggressive duplicate purging enabled (option -l 2) (Cheng et al. 2021). The primary contig assembly was used for all downstream analyses.

RNA was extracted from the head and silk gland, and library preparation was performed using the IsoSeq express
workflow. The library was then sequenced on a single Sequel II PacBio SMRT cell. Further details are provided in supplementary note 1, Supplementary Material online. The IsoSeq3 pipeline, part of the pbbioconda package, was used to generate IsoSeq clusters, following the published PacBio IsoSeq workflow (https://github.com/PacificBiosciences/IsoSeq/blob/master/isoseq-clustering.md, last accessed August 9, 2021). The steps in the pipeline are 1) circular consensus sequence calling (CCS read generation), 2) primer removal and demultiplexing, 3) refining (trimming of polyA tails and concatemer removal), 4) clustering, and 5) polishing.

We screened the genome assembly for potential contaminants with BlobTools v1.0 (Laetsch and Blaxter 2017) (supplementary note 2a and supplementary fig. 1, Supplementary Material online). We assessed genome quality and completeness with BUSCO v4.1.1 (Seppey et al. 2019) (supplementary note 3, Supplementary Material online) using the OrthoDB v.10 Insecta gene set (Kriventseva et al. 2019), and generated genome statistics using the assembly_stats v0.1.4 script (Trizna 2020) (supplementary table 1, Supplementary Material online, for full output). We conducted genome profiling (estimation of major genome characteristics such as size, heterozygosity, and repetitiveness) on the HiFi sequence data with GenomeScope 2.0 (Ranallo-Benavidez et al. 2020); these methods are described in supplementary note 4a and supplementary figures 3 and 4, Supplementary Material online.

Repeat and Gene Annotation

We identified and classified repetitive elements de novo and generated a library of consensus repeat sequences for the genome using RepeatModeler 2.0 (Flynn et al. 2020). We then annotated and masked repeats in the assembly with RepeatMasker 4.1.1 (Smit et al. 2013–2015) using the custom repeat library generated in the previous step. Finally, we reran RepeatMasker on the masked genome using the Repbase arthropod repeat library (Bao et al. 2015). We annotated the N. cornuta genome assembly using MAKER v3.01.03 (Cantarel et al. 2008) and generated ab initio gene predictions using SNAP (Korf 2004), with more details provided in supplementary note 5, Supplementary Material online. To generate functional predictions on the predicted proteins, we used Blast2GO. First, we extracted the CDS sequences from the genome and then used blastx (nr, e-value 1e-4, max_target_seqs = 5) to compare the predicted genes against the NCBI RefSeq nonredundant protein database. We used Blast2GO v1.4.4 (Götz et al. 2008) to map functional annotation and GO terms to the resulting sequences.

Supplementary Material

Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments

This research was supported by Grant-in-Aid for Scientific Research (KAKENHI) by Japan Society for the Promotion of Science (JSPS) (18H06077) to Y.I.; Smithsonian National Museum of Natural History Global Genome Initiative (GGI-Peer-2018-182) to T.P.C., R.D., T.D., A.Y.K., and P.B.F. We thank Donald Hobern and Martin Olofsson for providing photographs used in figure 1.

Author Contributions


### Table 1

<table>
<thead>
<tr>
<th>Order</th>
<th>Family</th>
<th>Species</th>
<th>Source</th>
<th>Accession</th>
<th>Assembly Length (Mb)</th>
<th>Contig N50 (Mb)</th>
<th>BUSCOs Present (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lepidoptera</td>
<td>Micropterigidae</td>
<td>Neomicropteryx cornuta</td>
<td>Present study</td>
<td>JAHKQU0000000000</td>
<td>541.12</td>
<td>16.92</td>
<td>98.1</td>
</tr>
<tr>
<td>Lepidoptera</td>
<td>Tineidae</td>
<td>Tinea trinotella</td>
<td>Wellcome Sanger Institute</td>
<td>GCA_905220615.1</td>
<td>371.74</td>
<td>13.66</td>
<td>98.2</td>
</tr>
<tr>
<td>Lepidoptera</td>
<td>Pieridae</td>
<td>Anthocaris cardamines</td>
<td>Wellcome Sanger Institute</td>
<td>GCA_905404175.1</td>
<td>359.62</td>
<td>6.27</td>
<td>98.9</td>
</tr>
<tr>
<td>Lepidoptera</td>
<td>Noctuidae</td>
<td>Autographa gamma</td>
<td>Wellcome Sanger Institute</td>
<td>GCA_905146925.1</td>
<td>373.07</td>
<td>11.19</td>
<td>99.1</td>
</tr>
<tr>
<td>Lepidoptera</td>
<td>Bombycidae</td>
<td>Bombyx mori</td>
<td>The University of Tokyo</td>
<td>GCA_014905235.2</td>
<td>460.35</td>
<td>12.20</td>
<td>98.6</td>
</tr>
<tr>
<td>Trichoptera</td>
<td>Limnephilidae</td>
<td>Hesperophylax magnus</td>
<td>Olsen et al. (2021)</td>
<td>JADDOG0000000000</td>
<td>1233.59</td>
<td>0.77</td>
<td>95.9</td>
</tr>
<tr>
<td>Trichoptera</td>
<td>Hydropsychidae</td>
<td>Hydropsyche teniens</td>
<td>Heckenhauer et al. (2019)</td>
<td>GCA_009617725.1</td>
<td>229.66</td>
<td>2.19</td>
<td>98.4</td>
</tr>
<tr>
<td>Trichoptera</td>
<td>Stenopsychidae</td>
<td>Stenopsyche tienmushanensis</td>
<td>Luo et al. (2018)</td>
<td>GCA_008973525.1</td>
<td>451.49</td>
<td>1.30</td>
<td>98.1</td>
</tr>
</tbody>
</table>

**Li et al.**


Downloaded from https://academic.oup.com/gbe/article/13/10/evab229/6380144 by guest on 17 November 2021

Data Availability
This project has been deposited at NCBI under the Bioproject ID: PRJNA731916 and the accession numbers: JAHKQU000000000 and SRR14882579. Annotations, predicted proteins, and other supplementary data are available on FigShare, https://doi.org/10.6084/m9.figshare.16775932.

Literature Cited
Associate editor: Adam Eyre-Walker

Downloaded from https://academic.oup.com/gbe/article/13/10/evab229/6380144 by guest on 17 November 2021