Supplementary Methods

Sizing up the onychophoran genome: repeats, introns, and gene family expansion contribute to genome gigantism in *Epiperipatus broadwayi*

Shoyo Sato, Tauana J. Cunha, Bruno A. S. de Medeiros, Danielle E. Khost, Timothy B. Sackton, Gonzalo Giribet

DNA Extraction

Epiperipatus broadwayi were collected in Tobago in 2017 and preserved in 95% ethanol. Specimens were deposited in the Invertebrate Zoology collection at the Museum of Comparative Zoology (MCZ) at Harvard University and can be accessed through the online database MCZbase (https://mczbase.mcz.harvard.edu/guid/MCZ:IZ:143930). Genomic DNA (gDNA) was extracted from trunk tissue of a single individual using a modified high salt protocol from the University of Liverpool [2]. The tRNA carrier was omitted from the protocol and a 45-minute-long incubation with RNase A at 37 °C was added to minimize RNA in the final extraction. The initial gDNA extraction was repaired with the PreCR Repair Mix (New England Biolabs) incubated at 37º C for 15 minutes followed by a chloroform cleanup modified from Dong [3] and Schalamun, Nagar [4]. These additional repair and cleanup steps were added following initial sequencing attempts which revealed evidence for poor Nanopore runs due to DNA impurity or damage. For the cleanup, 1 volume of chloroform:isoamyl alcohol (24:1) was added at the end of the PreCR incubation and mixed by inversion for 5 minutes followed by centrifuging at 20,000 g for 5 minutes, transferring the aqueous phase to a new tube and repeating the procedure for a total of two cleanups. Then 0.1 volume of 3 M sodium acetate (pH 5.3) and 1 volume of isopropanol chilled at 4º C were added. The tube was inverted a few times until the formation of a precipitate and then centrifuged at 20,000 g for 15 minutes at 4º C. After removing the supernatant, 1 mL 70% ethanol was added, centrifuged at 13,000 g for 5 minutes at 4º C and the

supernatant removed. This was repeated for a total of 2 washes. The tube was then quickly spun, and the remaining ethanol pipetted out, followed by air drying of the sample for additional 5 minutes. The DNA pellet was resuspended in 100 µL of 10 mM Tris-HCl pH 8 for 3 hours at room temperature.

Nanopore Sequencing

Following quality control, the resuspended DNA was divided into two aliquots. One aliquot was sheared by pipetting ten times with a p1000 pipette and cleaned using the Circulomics short read eliminator kit (Circulomics, Baltimore, MD, USA) and the other was directly cleaned using the Circulomics short read eliminator XL kit (Circulomics, Baltimore, MD, USA) as an attempt to obtain even longer reads. Separate library preparations were done on the aliquots using the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK). These followed manufacturer protocols with the following modifications: (1) DNA repair and end preparation with 30 minutes of incubation at 20º C and another 30 minutes at 65º C (2) incubation with magnetic beads during cleanup extended to 10 minutes and (3) elution following bead cleanup incubated for 20 minutes at 37º C. Subsequently, two initial sequencing runs using the MinION (Oxford Nanopore Technologies, Oxford, UK) were conducted. Flow cells were cleaned using the Flow Cell Wash Kit (Oxford Nanopore Technologies, Oxford, UK) and reloaded for a total of three loads per aliquot. Total time for sequencing was 72 h for each MinION run. After these initial runs, a larger sequencing run was conducted to increase coverage. DNA was extracted again from the same individual, using the same protocol, and was sheared to 29kb using Covaris g-tube (Covaris, Woburn, MA, USA) for 1 min at 6000 rpm in an Eppendorf 5415R centrifuge (Eppendorf, Hamburg, Germany). The library was prepared using

the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK) with the following modifications: 1) DNA repair and end preparation reaction was incubated for 10 min at 20 °C; 2) DNA was bead purified after ligation, washed with 2x LFB to enrich for larger fragments, and eluted for 10 min at 37 °C on a rotator. The finished library was loaded onto the PromethION (Oxford Nanopore Technologies, Oxford, UK) and ran for 72 h. This second round of library preparation and sequencing was conducted at the Bauer Core Facility at Harvard University.

Assembly

Unique 21mers were counted in the raw Illumina reads using Jellyfish [5]. The resulting histogram was plotted, and the genome size was estimated in R. Nanopore data were basecalled using Guppy v.4.5.2 [6] with disabled q-score filtering to maximize data input into assembly. Raw Hi-C reads were trimmed for adaptors and quality filtered using TrimGalore [7]. Raw Nanopore reads were assembled using Flye with default settings [8]. To polish the assembly, raw Nanopore reads were mapped against the draft genome using minimap2 [9] with the --map-ont preset, after which medaka (in batches of 1000 contigs to improve speed) was used to call consensus [10]. Then whole-genome Illumina reads were mapped against the medaka-polished draft using Bowtie2 [11] and used in single round of HyPo as a final polish [12]. Following assembly polishing, filtered and trimmed Hi-C reads were mapped to the genome using BWA [13] and the resulting BAM files were processed using the Arima-HiC mapping pipeline [14]. The resulting BED files were used to scaffold the initial contigs using SALSA2 [15, 16]. Scaffolds were screened for contaminants using Blobtools [17, 18] and the resulting scaffolds of dubious origin were manually checked using BLAST [19]. Haplotigs were removed from the filtered assembly using Purge Haplotigs [20]. The genome assembly workflow can be viewed in

figure 6a. Completeness of the final scaffolded assembly was assessed using BUSCO [21, 22] and the Arthropoda OrthoDB v10 [23].

Annotation

Transposable elements (TEs) were modeled and identified using RepeatModeler [24] using default settings followed by masking of the genome with RepeatMasker [25]. The masked assembly was then annotated with BRAKER2 [26-30] using three databases. A custom database was created using the transcriptomes sequenced in Baker, Buckman-Young [31]. Details of the transcriptomes including GenBank accession numbers and BUSCO scores can be found in supplementary table 2. The transcriptomes were translated to amino acids with TransDecoder v.5.5.0 [32] to accommodate the evolutionary distances between these sequences and the draft genome. The following peptide fastas were concatenated to make the onychophoran proteome database. Genes were also predicted using the Arthropoda OrthoDB v10 [23]. Additionally, the untranslated transcriptome of *Epiperipatus broadwayi* (GenBank: SRX10007847) from the same collection as our genome was used as a separate database for annotation. The three predictions from these separate databases were combined using TSEBRA [33] using low stringency settings to maximize potential gene predictions. Functional annotations were added to the final protein predictions using the online OMA browser [34] and eggNOG [35, 36]. The full annotation workflow can be viewed in figure 6b.

Figure 6. Bioinformatics pipelines used to assemble and annotate the genome of *Epiperipatus broadwayi*. (*a*) Pipeline used for the full assembly of the genome of *Epiperipatus broadwayi* including polishing, scaffolding, and filtering. (*b*) Pipeline used for the annotation of the genome of *Epiperipatus broadwayi* including data inputs.

Gene Expansion Analysis

Analysis of gene family expansion was conducted using Orthofinder [37]. The dataset included genomes of all panarthropod phyla with two tardigrades (*Hypbsibius dujardini, Ramazzottius varieornatus*) and nine genomes representing all major arthropod lineages (*Apis mellifera, Drosophila melanogaster, Trobolium castaneum, Folsomia candida, Daphnia pulex, Tigriopus californicus, Strigamia maritima, Stegodyphus mimosarum, Varroa destructor*) pulled from GenBank. Accession numbers of the genomes used in the analysis can be found in supplementary table 3. Due to the high gene count and the extensive expansion of several gene families, the predicted proteins were compared to the PFAM database [38] to check for the presence of transposable elements or transposable-element-associated proteins among annotated proteins. The entire amino acid fasta file was searched against PFAM using Hmmer v.3.2.1 [39] using a cutoff of 1e-6. The resulting Hmmer output was manually scanned to obtain a list of genes that hit to a TE associated domain. This list was compared to the genes in each orthogroup and taking a conservative approach, orthogroups consisting of at least 25% hits to the list were considered a TE or TE derived orthogroup. The combined annotations from TSEBRA were used to extract transcripts from the final assembly. Kallisto v0.46.1 [40] with 100 bootstrap samples was then used to quantify transcript abundances using RNA sequenced from trunk tissue. To check for the presence or absence of the putatively onychophoran specific orthogroups in other available genomes, a database was created from the *Epiperipatus broadwayi* amino acid fasta using DIAMOND [41]. Subsequently DIAMOND blastx [41] was run in sensitive mode with a cutoff of e-10 with the untranslated *Euperipatoides rowelli* genome from NCBI (GenBank:

GCA_003024985.2) and an additional newly sequenced genome of *Peripatoides* sp. (unpublished, https://mczbase.mcz.harvard.edu/guid/MCZ:IZ:144603).

Onychophoran Hemocyanins

To test the utility of this new genome, the genome was mined for hemocyanins, proposed to be a key factor in the evolution of Arthropoda. Putative hemocyanins were identified in the genome using two approaches. First a database of representative arthropod hemocyanin protein fastas were downloaded from GenBank (supplementary table 4) and used as 'baits' by querying against a custom BLAST database created from the *Epiperipatus braodwayi* proteome fasta with blastp v2.12.0 [19]. All hits with an *e*-value ≤ 1e-20 were stored. A custom HMM profile for hemocyanin was created using the same database of hemocyanin protein fastas as for the blast search and again used as 'baits' to search the *Epiperipatus broadwayi* proteome fasta with Hmmer v.3.2.1 [39]. All hits with an *e*-value ≤ 1e-30 were stored. Hits from both programs were filtered for redundancies and resulted in 31 unique transcripts from *Epiperipatus broadwayi* identified as putative hemocyanins. These transcripts were then used as queries for reciprocal BLAST searches against the NCBI nr database [42] using blastp v2.12.0 [19]. Storing the best hit for each transcript yielded 10 unique GenBank entries (supplementary table 4). All putative hemocyanin transcripts obtained from BLAST and Hmmer were combined with the initial representative hemocyanin sequences from NCBI and the reciprocal BLAST hits. Additional sequences of arthropod phenoloxidases were added as outgroups to the analysis (supplementary table 4) [43, 44]. A search for phenoloxidases was performed following the methods used for the hemocyanin search. This resulted in a redundant transcript set and was omitted. The resulting matrix was aligned using MAFFT [45]. Gaps, divergent sequences, and short sequences were

removed from the multiple sequence alignment using CIAlign v1.0.17 [46]. Model testing, phylogenetic tree reconstruction, and branch support assessment was conducted using IQ-TREE [47-49].

References

1. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Research, 2016. **44**(D1): p. D733–D745.
2. Donnan Laboratories, *Extraction of DNA from tissue: high salt method. Version 1.0*. 2001: School of Biological Sciences, University of Liverpool, L69 7ZD, UK.
3. Dong, C., *Purification of HMW DNA from Fungi for long read sequencing*. 2017: protocols.io.
4. Schalamun, M., et al., *Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora.* Molecular Ecology Resources, 2019. **19**(1): p. 77–89.
5. Marcais, G. and C. Kingsford, *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.* Bioinformatics, 2011. **27**(6): p. 764–770.
6. Oxford Nanopore Technologies, *Guppy*. 2021.
7. Krueger, F., *Trim Galore!* 2021.
8. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nature Biotechnology, 2019. **37**(5): p. 540–546.
9. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094–3100.
10. Oxford Nanopore Technologies, *medaka*. 2018.
11. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, 2012. **9**(4): p. 357–359.
12. Kundu, R., J. Casey, and W. Sung, *HyPo: Super fast & accurate polisher for long read genome assemblies.* bioRxiv, 2019.
13. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754–1760.
14. ArimaGenomics. *Arima-HiC Mapping Pipeline*. 2019; Available from: https://github.com/ArimaGenomics/mapping_pipeline.
15. Ghurye, J., et al., *Scaffolding of long read assemblies using long range contact information.* BMC Genomics, 2017. **18**(1): p. 527.
16. Ghurye, J., et al., *Integrating Hi-C links with assembly graphs for chromosome-scale assembly.* PLoS Computational Biology, 2019. **15**(8): p. e1007273.
17. Laetsch, D.R. and M.L. Blaxter, *BlobTools: Interrogation of genome assemblies.* F1000Research, 2017. **6**.
18. Laetsch, D.R., et al., *DRL/blobtools: BlobTools*. 2017.
19. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**: p. 403–410.
20. Roach, M.J., S.A. Schmidt, and A.R. Borneman, *Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies.* BMC Bioinformatics, 2018. **19**(1): p. 460.
21. Simão, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210–3212.
22. Waterhouse, R.M., et al., *BUSCO applications from quality assessments to gene prediction and phylogenomics.* Molecular Biology and Evolution, 2018. **35**(3): p. 543–548.
23. Kriventseva, E.V., et al., *OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.* Nucleic Acids Research, 2019. **47**(D1): p. D807–D811.
24. Flynn, J.M., et al., *RepeatModeler2 for automated genomic discovery of transposable element families.* Proceedings of the National Academy of Sciences of the USA, 2020. **117**(17): p. 9451–9457.

25. Smit, A.F.A., S. Hubley, and P. Green, *RepeatMasker*. 2021.
26. Bruna, T., et al., *BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.* NAR Genomics Bioinformatics, 2021. **3**(1): p. lqaa108.
27. Hoff, K.J., et al., *BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS.* Bioinformatics, 2016. **32**(5): p. 767–769.
28. Hoff, K.J., et al., *Whole-genome annotation with BRAKER.* Methods in Molecular Biology, 2019. **1962**: p. 65–95.
29. Stanke, M., et al., *Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.* BMC Bioinformatics, 2006. **7**: p. 62.
30. Stanke, M., et al., *Using native and syntenically mapped cDNA alignments to improve de novo gene finding.* Bioinformatics, 2008. **24**(5): p. 637–644.
31. Baker, C.M., et al., *Phylogenomic analysis of velvet worms (Onychophora) uncovers an evolutionary radiation in the Neotropics.* Molecular Biology and Evolution, 2021. **38**(12): p. 5391–5404.
32. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nature Protocols, 2013. **8**(8): p. 1494–1512.
33. Gabriel, L., et al., *TSEBRA: transcript selector for BRAKER.* BMC Bioinformatics, 2021. **22**(1): p. 566.
34. Altenhoff, A.M., et al., *The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements.* Nucleic Acids Research, 2015. **43**(Database issue): p. D240–D249.
35. Cantalapiedra, C.P., et al., *eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale.* Molecular Biology and Evolution, 2021. **38**(12): p. 5825–5829.
36. 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, 2019. **47**(D1): p. D309–D314.
37. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for comparative genomics.* Genome Biology, 2019. **20**(1): p. 238.
38. Mistry, J., et al., *Pfam: The protein families database in 2021.* Nucleic Acids Research, 2021. **49**(D1): p. D412–D419.
39. Eddy, S.R., *Accelerated profile HMM searches.* PLos Computational Biology, 2011. **7**(10): p. e1002195.
40. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nature Biotechnology, 2016. **34**(5): p. 525–527.
41. Buchfink, B., K. Reuter, and H.G. Drost, *Sensitive protein alignments at tree-of-life scale using DIAMOND.* Nature Methods, 2021. **18**(4): p. 366–368.
42. Sayers, E.W., et al., *Database resources of the national center for biotechnology information.* Nucleic Acids Research, 2022. **50**(D1): p. D20–D26.
43. Burmester, T., *Molecular evolution of the arthropod hemocyanin superfamily.* Molecular Biology and Evolution, 2001. **18**(2): p. 184–195.
44. Kusche, K., H. Ruhberg, and T. Burmester, *A hemocyanin from the Onychophora and the emergence of respiratory proteins.* Proceedings of the National Academy of Sciences of the USA, 2002. **99**(16): p. 10545–10548.
45. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Molecular Biology and Evolution, 2013. **30**(4): p. 772–780.

46. Tumescheit, C., A.E. Firth, and K. Brown, *CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments.* PeerJ, 2022. **10**: p. e12983.
47. Nguyen, L.T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.* Molecular Biology and Evolution, 2015. **32**(1): p. 268–274.
48. Hoang, D.T., et al., *UFBoot2: Improving the ultrafast bootstrap approximation.* Molecular Biology and Evolution, 2017. **35**(2): p. 518–522ka.
49. Kalyaanamoorthy, S., et al., *ModelFinder: fast model selection for accurate phylogenetic estimates.* Nature Methods, 2017. **14**(6): p. 587–589.