Version 2.0, Updated 2023.05.24

Checklist for Reproducible Publishing of 'Omics in Emerging Model Organisms

Date initiated: 1/1/18

Date completed: 12/23/20

Date(s) of revision: 12/23/20

Authors: Varney RM, Speiser DI, McDougall C, Degnan BM, Kocot KM

Title of study: The Iron-Responsive Genome of the Chiton Acanthopleura granulata

A. RATIONALE

1. Goal of present study ..... ☑
2. What products are/will be produced? (e.g. draft genome, annotation, single-cell seq) ..... ☑

Goal of study is a high-quality genome and annotation for the chiton Acanthopleura granulata. Products will be a genome and annotation, including several transcriptomes from various tissues across the same chiton, as well as transcriptomes from other species of chitons for annotation purposes.

B. SPECIMEN AND TISSUE INFORMATION

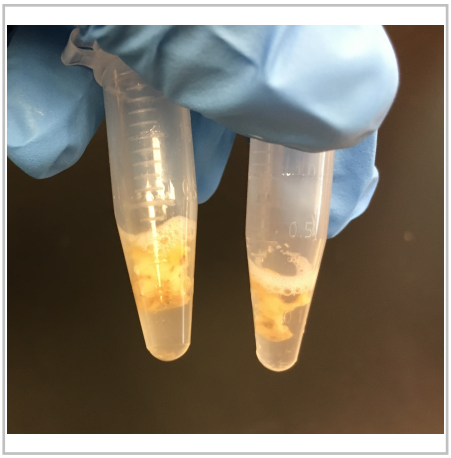
3. Resources for specimen identification acknowledged ..... ☑
4. Collection information (where/when/how), and relevant permit information ..... ☑
   a. Images of collection locations included (see section H, voucher information) ..... ☑
5. Dissection descriptions if relevant at this stage ..... ☑
   a. Images of dissection included ..... ☑
6. Storage information (duration, temperature, transport, etc.) ..... ☑

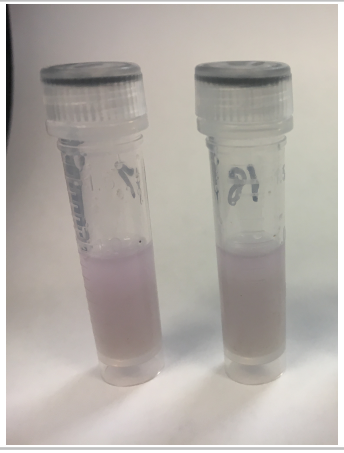
Specimens were ID'ed via Kaas and Van Belle Monograph of Living Chitons, vol. 2. Specimen was collected from Harry Harris State Park in the Florida Keys under Special Activity License #SAL-17-1983-SR, coordinates approximately 25.025056, -80.493593. Specimens were brought back to the lab live in a bucket with a single aerator and air conditioning running the entire drive. Water was changed halfway back (approx. 4 out of 8 hours). Note that a lid was vital to retain specimens inside bucket.

C. EXTRACTION INFORMATION

7. Method of extraction ☑
8. Citation for extraction method (cite in any manuscripts as well!) ☑
9. Rationale for extraction method (included notes about any methods attempted that failed, and especially any modifications made to the method used.) ☑
10. Number of individuals (if pooled) ☑
11. Tissue(s) used ☑
12. Amounts of tissue (and/or photo of tissue sample with scale) ☑

We extracted DNA with a customized CTAB-phenol chloroform method (full protocol available on Dryad associated with manuscript) from ~4mm2 chunks of muscle tissue (foot). DNA for short reads was cleaned with a Zymo Clean and Concentrator Kit. For transcriptomes, we extracted RNA with the Omega Bio-tek EZNA Mollusc RNA Kit following manufacturer's instructions from girdle, ctenidia, gonad, foot, and four regions of the radula. Note that the iron caused a purple color in radula samples (picture below); this did not appear to impact extractions negatively.





D. SEQUENCING INFORMATION

13. Sequencing technology(ies) used ☑
14. Library preparation details ☑
    a. If a core facility was used, which core? ☑
15. Flowcell details ☑
16. Data collection information (e.g. movie time, settings) ☑
17. Basecaller and version ☑
18. Raw data storage after sequencing ☑

For genome sequencing, we used a Nanopore GridION. Two sequencing libraries were prepared with Nanopore ligation kit LSK-108 and sequenced these on 2 FloMin106-R9.4.1 flow cells. Then we prepared two additional libraries with updated kit LSK-109 and sequenced these on 2 R9.4.1 RevD flowcells. Each run was permitted to run for a full 48 hours. We basecalled with Guppy 4.0. Raw data was stored immediately on our local RAID drive until uploading to NCBI. For transcriptomes, cDNA was synthesized and complementary DNA (cDNA) amplied from each tissue using the SmartSeq v4 Ultra Low-input RNA kit (Clontech) from 1 ng of input RNA with 17 cycles of PCR. Eight dual-indexed sequencing libraries were created with the Illumina Nextera XT kit, using 1 ng of input cDNA. We sent the eight libraries to Macrogen (Seoul, South Korea) where they were pooled and sequenced on one lane of an Illumina HiSeq 4000 (2 × 100 bp paired-end reads). Raw data was stored immediately on our local RAID drive until uploading to NCBI.

E. ASSEMBLY INFORMATION

19. Assembler(s) and version(s) ..... ☑
20. Description of adjustments for heterozygosity ..... ☑
21. Scaffolding technology choice and rationale ..... ☑
22. Scaffolding library preparation and sequencing details ..... ☑
23. Scaffolding software and version ..... ☑
24. Decontamination software and version ..... ☑
    a. Location of contaminant sequences if available ..... ☑
25. Final assembly information:
    a. Total number of all contigs ..... ☑
    b. Total number of scaffolds ..... ☑
    c. Number of contigs/scaffolds included in subsequent analyses ..... ☑
26. All code is available ..... ☑

The genome was first assembled with MaSuRCA v. 3.3.5, producing an assembly with 2,858 contigs. Next, Redundans v. 0.14a was used to collapse heterozygosity, decreasing the assembly to 1,285 contigs. Blobtools2 was used to decontaminate the genome. One lane of Bionano SAPHYR optical mapping was used to scaffold the genome, using enzymes BssSI and DLE1, resulting in 87 scaffolds. All scaffolds were included in subsequent analyses, and all code is available in the supplemental document.

F. ANNOTATION APPROACH

27 Assembly selected for annotation ..... ☑
28. Annotation pipeline selection, version, and rationale ..... ☑
29. Evidence files used from the present study ..... ☑
30. Evidence files used from previous studies (citations included in manuscript) ..... ☑
31. All code is available ..... ☑

Annotation was performed on the 87 contig assembly. We initially attempted to use PASA/EVM for annotation, but found that EVM reduced the number of "passing" gene models to <5000 each try. This is believed to be a result of a high degree of alternative splicing, especially including alternative splicing of exon 1. Instead, we first ran MAKER 2.31.10 with a composite transcriptome of all the tissues we extracted RNA from. We used the gene models with AED<0.5 to train Augustus v3.0.3. Separately, we ran PASA 2.4.1 with the composite transcriptome as evidence. Instead of EVM, we used the intersect tool from BEDtools v.2.29.2 to remove identical sequences. We then clustered this set of gene models (81,691) with cdhit-EST v. 4.8.1 run with the -g flag and a cluster threshold of 0.8. This produced a set of 20,470 genes. All code and commands are available in the supplemental document of the manuscript.

G. DATA AVAILABILITY

32. Raw data accesstion number(s) listed ..... ☑
33. Assembly accession number(s) listed if applicable ..... ☐
34. Annotation is uploaded to NCBI/EBI/DDBJ ..... ☑
    AND/OR
35. Annotation is available on an online repository (e.g. Dryad, Figshare) ..... ☑
36. All manuscript files are available for reviewers to verify intent to share all data contained within the published manuscript (e.g. annotations, assemblies, metadata) ..... ☑

All sequences were uploaded to NCBI BioProject PRJNA578131. Annotations are available on dryad, https://doi.org/10.5061/dryad.wstqjq2k9. Files for review purposes are shared in a Box folder, to be included in initial submissions as a reviewer resource.

H. VOUCHER AVAILABILITY

37. A voucher is available in a museum for the *ome(s) presented in the current manuscript ☑
38. Photographs are available for the *ome(s) presented in the current manuscript ..... ☑
39. A voucher is available for a different specimen collected at the same place/time ..... ☑

Remaining tissue from the specimen used for genome sequencing is deposited in the Alabama Museum of Natural History Invertebrate Zoology collection. Additional specimens are available from the laboratory collections of KM Kocot (KK702.704).

40. Museum information is in the manuscript ..... ☑

AMNH 21276

I. OTHER NOTES

I should have taken photographs of the actual dissections.