Metagenomic data for Halichondria panicea from Illumina and nanopore sequencing and preliminary genome assemblies for the sponge and two microbial symbionts

Strehlow, Brian W.; Schuster, Astrid; Francis, Warren R.; Canfield, Donald E.

Published in:
BMC Research notes

DOI:
10.1186/s13104-022-06013-3

Publication date:
2022

Document version:
Final published version

BMC Research Notes

DATA NOTE

Open Access

Metagenomic data for *Halichondria panicea* from Illumina and nanopore sequencing and preliminary genome assemblies for the sponge and two microbial symbionts

Brian W. Strehlow*, Astrid Schuster, Warren R. Francis and Donald E. Canfield

Abstract

Objectives: These data were collected to generate a novel reference metagenome for the sponge *Halichondria panicea* and its microbiome for subsequent differential expression analyses.

Data description: These data include raw sequences from four separate sequencing runs of the metagenome of a single individual of *Halichondria panicea*—one Illumina MiSeq (2 × 300 bp, paired-end) run and three Oxford Nanopore Technologies (ONT) long-read sequencing runs, generating 53.8 and 7.42 Gbp respectively. Comparing assemblies of Illumina, ONT and an Illumina-ONT hybrid revealed the hybrid to be the 'best' assembly, comprising 163 Mbp in 63,555 scaffolds (N50: 3084). This assembly, however, was still highly fragmented and only contained 52% of core metazoan genes (with 77.9% partial genes), so it was also not complete. However, this sponge is an emerging model species for field and laboratory work, and there is considerable interest in genomic sequencing of this species. Although the resultant assemblies from the data presented here are suboptimal, this data note can inform future studies by providing an estimated genome size and coverage requirements for future sequencing, sharing additional data to potentially improve other suboptimal assemblies of this species, and outlining potential limitations and pitfalls of the combined Illumina and ONT approach to novel genome sequencing.

Keywords: Metagenome, Hologenome, *Halichondria panicea*, Porifera, Microbiome, Objective

Objective

These data were generated to create a reference metagenome for the emerging model sponge species, *Halichondria panicea* and its microbiome. The goal was then to use this reference to study changes in gene expression under different oxygen concentrations in order to understand how this species tolerates hypoxia [see 1]. During the process of data collection, Knobloch et al. [2] generated a reference genome for the dominant microbial symbiont 'Candidatus Halichondribacter symbioticus', and the data presented here were not sufficient to construct a suitable reference genome for the sponge, limiting the scope of these data for a full research paper.

Given the considerable interest in *H. panicea* and its widespread distribution, we think that the data provided can inform future experiments and contribute to a more complete genome later. Finally, by sharing suboptimal data we aimed to identify some potential pitfalls for future genome projects, particularly those of poriferans.

*Correspondence: strehlow@biology.sdu.dk
Department of Biology & Nordcee, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark

Data description

Sample collection and DNA extraction

To limit assembly issues caused by allelic variation, a single individual of *H. panicea* (approximately 1 g of

tissue [wet weight]) was collected from the side of a pier manually (while wearing gloves) within the inlet to Kerteminde Fjord in Denmark (decimal degrees: 55.449808, 10.661299) in 2018. The tissue was immediately cut on a sterile surface with a sterile scalpel, placed into sterile 1.5 mL cryovials, and flash frozen in liquid nitrogen. DNA was extracted and purified using a modified phenol–chloroform extraction (see [3] for full protocol) under sterile laboratory conditions. Microbes were not physically separated from sponge tissue before DNA extractions or sequencing. This protocol yielded the highest quality DNA and highest concentrations above 15,000 bp compared to five different extraction protocols (see supplemental material in [4]).

In total, nine micrograms of double stranded DNA were extracted and Nanodrop A260/A280 and A260/A230 ratios were 1.79 and 2.17, respectively. The DNA integrity number (DIN) was 1.6, with high concentrations of DNA between 100 and 4000 base pairs (bp). A smearing pattern in gels was observed for all DNA extractions of *H. panicea* using various protocols (see supplementary material in [4]). This pattern could indicate high levels of degradation; however, a substantial amount of DNA was still intact and > 15,000 bp long in samples used for sequencing.

Sequencing

Approximately 1 µg of DNA was sequenced on an Illumina MiSeq sequencer (2 × 300 bp, paired-end, Illumina, Inc). This run generated 356 million paired-end reads (53.8 Gbp).

The first sequencing run using Oxford Nanopore Technologies (ONT) generated 1.26 million reads (3.4 Gbp, read N50: 2700 bp, longest read: 39,702 bp). For more details on the sequencing methods, see the supplemental material in [4].

Due to a low coverage of Opisthokonta contigs (from the Illumina data) in the nanopore reads, two additional rounds of nanopore sequencing were performed after whole genome amplifications (WGA, see supplementary material), generating 4.021 Gbp from the amplified *H. panicea* DNA. A summary of the public locations of all data generated is shown in Table 1.

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	DNA extraction protocol	.io	Protocols.io, https://doi.org/10.17504/protocols.io.yvkfw4w [3]
Data set 1	Illumina raw sequences lane 1	fastq	NCBI Sequence Read Archive, https://identifiers.org/insdc.sra:SRR15711138 [5]
Data set 2	Illumina raw sequences lane 2	fastq	NCBI Sequence Read Archive, https://identifiers.org/insdc.sra:SRR15711137 [6]
Data set 3	Nanopore run 1 raw sequences	fastq	NCBI Sequence Read Archive, https://identifiers.org/insdc.sra:SRR15711136 [7]
Data set 4	Nanopore run 2 WGA raw sequences	fastq	NCBI Sequence Read Archive, https://identifiers.org/insdc.sra:SRR15711135 [8]
Data set 5	Nanopore run 3—WGA raw sequences	fastq	NCBI Sequence Read Archive, https://identifiers.org/insdc.sra:SRR15711134 [9]
Data set 6	Whole metagenome assembly (from Illumina sequences)	fasta	NCBI Assembly, https://identifiers.org/assembly:GCA_020423325.1 [10]
Data set 7	*H. panacea* genome assembly (bin 1 from Illumina sequences)	fasta	NCBI Assembly, https://identifiers.org/assembly:GCA_020423275.1 [11]
Data set 8	HOC36 bin assembly (from Illumina sequences)	fasta	NCBI Assembly, https://www.ncbi.nlm.nih.gov/assembly/GCA_020423265.1 [12]
Data set 9	Proteobacteria bin assembly (from Illumina sequences)	fasta	NCBI Assembly, https://identifiers.org/assembly:GCA_020423255.1 [13]
Data set 10	Nanopore only metagenome assembly	fasta	NCBI Assembly, https://www.ncbi.nlm.nih.gov/assembly/GCA_020423245.1 [14]
Data set 11	Hybrid nanopore-Illumina assembly	fasta	NCBI Assembly, https://identifiers.org/assembly:GCA_020423345.1 [15]
Data file 2	Supplementary material	pdf	Harvard Dataverse, https://doi.org/10.7910/DVN/DJYOOI [4]

Genome assembly and annotation

Illumina metagenome assembly

Full details of quality control, binning, assembly and annotation of the metagenome are in the supplementary material. Three bins were produced including: (1) a large Opisthokonta bin, which was labeled as the sponge bin [11]; (2) a bin for a *Gammaproteobacteria* of the order 'HOC46' [12]; and (3) a *Proteobacteria* bin ([13], Table 1). The sponge bin was highly fragmented (63,555 scaffolds) and contained only 51.57% of core metazoan genes (with 77.46% partial matches, Supplemental Table 1 in [4]) measured using BUSCOv5 [16]. More bins could potentially be extracted from these data in the future.

The two bacterial genome bins were annotated using PROKKA v. 1.14 [17], and their completeness was estimated with CheckM [18] (see supplemental in [4] for more information about these two bins).

ONT and hybrid assemblies

Two additional metagenome assemblies were made using (1) ONT data from all three sequencing runs and (2) a combination of Illumina and ONT data. The second ONT sequencing run (following WGA) had high percentages of contamination (8%) and chimerism (5–10%). These ONT data were polished and filtered to remove these errors as described in the supplementary material. A summary of the nanopore-only metagenome assembly is shown in Supplemental Table 2 [4].

The ONT and Illumina (Supplemental Table 1 in [4]) sponge assemblies were merged to create a hybrid metagenome using Flye v.2.6 [19] (Supplemental Table 2 in [4]).

Limitations

Although the incorporation of long read nanopore data in the hybrid assembly did slightly increase the metagenome N50 and decrease the number of scaffolds in the assembly, the genome was still highly fragmented. A major limitation in sponge genomics that is often discussed but rarely written about is the difficulty in extracting high quality, high molecular weight DNA. This difficulty was likely either a result of some innate, highly efficient DNA degradation pathway in *H. panicea* or indicated the presence of DNA and/or degradation pathways from associated microorganisms or secondary metabolites. Obtaining high molecular weight DNA is paramount for successful long-read sequencing as well as genome assembly downstream regardless of sequencing technique. ONT sequencing can selectively sequence smaller DNA fragments if they are present. Additionally, microbial diversity within the metagenome and potential genetic variation caused by diploidy could also have limited genomic assembly.

This note represents the first attempt to sequence a sponge genome using Nanopore and Illumina sequencing, so improved genomic DNA recovery might validate this combination of methods, although it is unclear how DNA recovery could be improved. However, at least 9,000 Mbp long reads need to be generated. Similarly, the coverage of ONT reads would need to be increased to ~70× to permit a better assembly. Additionally, WGA should be used with caution due to the high rates of chimerism and contamination throughout the process. Improving coverage would also improve the assembly of prokaryotic genomes in the metagenome.

Recently, the generation of a near-chromosome level scaffolded genome assembly for the sponge *Ephydatia muelleri* was accomplished using PacBio, Chicago, and Dovetail Hi-C libraries sequenced to ~1490× coverage [20]. This sequencing method may therefore be the best for de novo genomes. The use of a sponge with limited microbial 'contamination' might also be critical for smooth genome assembly, although this effectively limits metagenomic projects. Finally, the use of a single haploid cell, like a sperm or egg cell, could improve future genome assembly performance by limiting allelic variation. However, single cell genomics could be limited by the amount and quality of DNA that can be isolated from a single cell.

Abbreviations

bp: Base pair(s); ONT: Oxford Nanopore Technologies; WGA: Whole genome amplification.

Acknowledgements

Special thanks to DNASense and Rasmus Dam Wollenberg for the sequencing and downstream support.

Author contributions

All authors participated in the conception and planning of the project and reviewed and contributed to drafts of the paper. BWS collected and analyzed the data, contributed reagents/materials/analysis tools, prepared the first draft of the paper, and prepared figures and tables. AS collected the data and contributed reagents/materials/analysis tools. WRF analyzed the data. DEC contributed reagents/materials/analysis tools. All authors read and approved the final manuscript.

Funding

This project was funded by Villum Fonden (Grant No. 16518).

Availability of data and materials

The data described in this Data Note can be freely and openly accessed on NCBI under the project https://identifiers.org/ncbi/bioproject:PRJNA753045. Each individual dataset [5–15], the DNA extraction protocol [3], and the supplementary material [4] can be accessed through links in Table 1 and the "References" section.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they do not have competing interests.

Received: 3 November 2021 Accepted: 23 March 2022
Published online: 09 April 2022

References

1. Mills DB, Ward LM, Jones C, Sweeten B, Forth M, Treusch AH, et al. Oxygen requirements of the earliest animals. Proc Natl Acad Sci. 2014;111:4168–72. https://doi.org/10.1073/pnas.1400547111.
2. Knobloch S, Jóhannsson R, Marteinsson VÞ. Genome analysis of sponge symbiont 'Candidatus Halichondribacter symbioticus' shows genomic adaptation to a host-dependent lifestyle. Environ Microbiol. 2020;22:483–98.
3. Wollenberg RD, Strehlow BW, Schuster A. Extracting high molecular weight DNA from *Halichondria panicea* (Phylum: Porifera). protocols.io. 2021. https://www.protocols.io/view/extracting-high-molecular-weight-dna-from-halichon-yvkfw4w.
4. Strehlow BW, Schuster A, Francis WR, Canfield DE. Supplementary material for: metagenomic data for *Halichondria panicea* from Illumina and Nanopore sequencing and preliminary genome assemblies for the sponge and two microbial symbionts. 2021. Harvard Dataverse. https://doi.org/10.7910/DVN/DJYOOI.
5. Strehlow BW, Schuster A, Francis WR, Canfield DE. Illumina raw sequences lane 1. NCBI. 2021. https://identifiers.org/insdc.sra:SRR15711138.
6. Strehlow BW, Schuster A, Francis WR, Canfield DE. Illumina raw sequences lane 2. NCBI. 2021. https://identifiers.org/insdc.sra:SRR15711137.
7. Strehlow BW, Schuster A, Francis WR, Canfield DE. Nanopore run 1 raw sequences. NCBI. 2021. https://identifiers.org/insdc.sra:SRR15711136.
8. Strehlow BW, Schuster A, Francis WR, Canfield DE. Nanopore run 2 WGA raw sequences. NCBI. 2021. https://identifiers.org/insdc.sra:SRR15711135.
9. Strehlow BW, Schuster A, Francis WR, Canfield DE. Nanopore run 3—WGA raw sequences. NCBI. 2021. https://identifiers.org/insdc.sra:SRR15711134.
10. Strehlow BW, Schuster A, Francis WR, Canfield DE. Whole metagenome assembly (from Illumina sequences). NCBI. 2021. https://identifiers.org/assembly:GCA_020423325.1.
11. Strehlow BW, Schuster A, Francis WR, Canfield DE. *H. panacea* genome assembly (bin 1 from Illumina sequences). NCBI. 2021. https://identifiers.org/assembly:GCA_020423275.1.
12. Strehlow BW, Schuster A, Francis WR, Canfield DE. HOC36 bin assembly (from Illumina sequences). NCBI. 2021. https://www.ncbi.nlm.nih.gov/assembly/GCA_020423265.1.
13. Strehlow BW, Schuster A, Francis WR, Canfield DE. Proteobacteria bin assembly (from Illumina sequences). NCBI. 2021. https://identifiers.org/assembly:GCA_020423255.1.
14. Strehlow BW, Schuster A, Francis WR, Canfield DE. Nanopore only metagenome assembly. NCBI. 2021. https://www.ncbi.nlm.nih.gov/assembly/GCA_020423245.1.
15. Strehlow BW, Schuster A, Francis WR, Canfield DE. Hybrid nanopore-Illumina assembly. NCBI. 2021. https://identifiers.org/assembly:GCA_020423345.1.
16. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
17. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
18. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
19. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.
20. Kenny NJ, Francis WR, Rivera-Vicéns RE, Juravel K, de Mendoza A, Díez-Vives C, et al. Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*. Nat Commun. 2020;11:1–11.

Publisher's Note