Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

Evaluation of *de novo* assembly technique in the South African abalone *Haliotis midae* transcriptome: A comparison from Illumina and 454 systems

Barbara Picone *, Clint Rhode, Rouvay Roodt-Wilding

Department of Genetics, Stellenbosch University, van der Bijl Street, JC Smuts Building, Private Bag X1, Matieland 7602, South Africa

ARTICLE INFO

Article history:
Received 24 October 2016
Received in revised form 4 November 2016
Accepted 9 November 2016
Available online 13 November 2016

Keywords:
Abalone
Haliotis midae
Hybrid assembly
Transcriptome

ABSTRACT

Next generation sequencing platforms have recently been used to rapidly characterize transcriptome sequences from a number of non-model organisms. The present study compares two of the most frequently used platforms, the Roche 454-pyrosequencing and the Illumina sequencing-by-synthesis (SBS), on the same RNA sample obtained from an intertidal gastropod mollusc species, *Haliotis midae*. All the sequencing reads were deposited in the Short Read Archive (SRA) database are retrievable under the accession number [SRR071314 (Illumina Genome Analyzer II)] and [SRR1737738, SRR1737737, SRR1737735, SRR1737734 (454 GS FLX)] in the SRA database of NCBI. Three transcriptomes, composed of either pure 454 or Illumina reads or a mixture of read types (Hybrid), were assembled using CLC Genomics Workbench software. Illumina assemblies performed the best *de novo* transcriptome characterization in terms of contig length, whereas the 454 assemblies tended to improve the complete assembly of gene transcripts. Both the Hybrid and Illumina assemblies produced longer contigs covering more of the transcriptome than 454 assemblies. However, the addition of 454 significantly increased the number of genes annotated.

Specifications	
Organism/cell line/tissue	*Haliotis midae*/muscle; ganglion, hepato-pancreas, gonad and gill
Sex	Three males and Three females
Sequencer or array type	Illumina Genome Analyzer II and 454 GS FLX platform
Data format	Raw data in FastaQ format
Experimental factors	Transcriptome profiling of four replicons of *Haliotis midae*
Experimental features	Specimens were collected across the geographical range of *H. midae* and biopsied tissues were immediately placed in RNALater® (Ambion®) solution and stored at −20 °C until RNA extraction could be performed.
Consent	NA
Sample source location	Saldanha Bay; Witsand; Riet Point, South Africa (33°02′ 40.64″ S; 17°56′00.53″E; 34°20′53.37″S; 19°01′39.75″E; 33°31′29.31″ S; 27°06′51.18″E.

1. Direct link to deposited data

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA79815 for *Haliotis midae* (Illumina);

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA257776 for *Haliotis midae* (454);

2. Introduction

The South African abalone (*Haliotis midae*) is a marine gastropod that has a large geographic distribution from west coast (Saldahna) to east coast (Riet Point). Therefore, the environmental conditions that this species is exposed to varies, making this organism a good candidate to explore the genetics and gene expression changes, which allow it to persist in such a dynamic environment. Although 454 is appropriate for assembly, the millions of short reads produced by Illumina provides a good reference for *de novo* transcriptome characterization in terms of contig length, transcriptome coverage, and complete assembly of gene transcripts [6]. Recent work comparing technologies suggest that hybrid assemblies combining 454 and Illumina reads yield the highest quality transcriptomes [3]. The present study used and combined Illumina and 454 pyrosequencing data, previously published [2,4], to

* Corresponding author.
E-mail addresses: barbara.picone@gmail.com (B. Picone), clintr@sun.ac.za (C. Rhode), roodt@sun.ac.za (R. Roodt-Wilding).

http://dx.doi.org/10.1016/j.gdata.2016.11.006

Table 1
Summary of assembly statistic.

Assembly	Number of contigs	N50	Average length (bp)	Maximum contig length	Total read length (total bases)
Illumina	22,761	501	260	10,744	10,635,178
454	31,491	395	393	4915	12,152,804
Hybrid	41,106	530	390	7193	11,404,294

characterize genes from the abalone *Haliotis midae*. Using Illumina sequencing on *H. midae*, Franchini et al. [2] were able to create a large EST library with over 1.1 billion bases suitable for *de novo* transcriptome assembly and gene/annotation analysis. In the view of the short read length, Picone et al. [4] used 454 pyrosequencing (≈450bp read length) to help implement and recognize predicting genes that might have been missed by the Illumina sequencing study. This report compares the success of these two technologies in identifying or predicting genes which are important for adaptation across broad environmental changes and for stress response to micro geographic environmental fluctuations. Furthermore, a hybrid assembly was constructed by merging contigs from the 454 and Illumina data.

3. Experimental design, materials and methods

3.1. Sampling preparation and assembly

Sequence processing and transcriptome assembly were conducted as previously described by Franchini et al. [2]. To obtain a comprehensive transcriptome, mRNA from various tissues was sequenced using either Roche 454 pyrosequencing or Illumina SBS (sequencing-by-synthesis). The extraction protocol as described in Van der Merwe [5] was used for the 454 samples. A protocol for extraction adapted from Falcao et al. [1] was followed for the Illumina samples. Illumina sequencing was performed using four-color DNA SBS technology; 454 pyrosequencing was conducted on a full-plate of the 454 GS FLX platform (Roche). CLC Genomics Workbench (8.5) as used to perform the individual as well as the hybrid assemblies.

3.2. Raw reads, base pairs and assemblies

The results of the Illumina and 454 sequencing for the different abalone samples are shown in Table 1. One of the goals of the present study was to perform a hybrid assembly to detect improvements over the individual assemblies. The hybrid assembly displayed the highest values across most standard metrics of transcriptome assembly (Table 1), followed by the Illumina assembly. The distribution of the contig lengths was quite different between the 454- and Illumina datasets. The Illumina and hybrid assemblies generated many contigs that were long; in contrast, the 454 assembly tended to have short contigs (*e.g.*, N50, longest contigs) and together with the hybrid assembly covered a large portion of the transcriptome. The absolute longest contigs were derived from the Illumina assembly (Table 1).

Table 2
Summary of gene ontology annotation.

Sequencer	Biological processes	Molecular function	Cellular component
454	52.9%	20.6%	37.3%
Illumina	42.1%	33.2%	24.7%

3.3. Annotation, orthology and gene ontology

De novo assembled transcriptomes from non-model species rely on BLAST-based annotations to provide information about gene identity and function. BLAST matching of the transcripts files (both Illumina and 454) showed that almost all the top hits matched sequences from other distant taxa (data not shown). For the Illumina samples, 3841 out of 22,761 contigs were annotated by BLAST and 5969 matched known proteins (*E*-value threshold of 10^{-3}). The number of sequences annotated by BLAST for 454 samples was higher (20,275 out of 31,491) and a total of 11,675 contigs matched known proteins (*E*-value threshold of 10^{-3}). Overall, the 454 reads contributed an additional 5706 annotations over the Illumina assembly, whereas the Illumina data added 2128 gene annotations over the 454 transcriptome. In order to identify orthologs and remove any redundancy, contigs from both datasets were subject to a reciprocal BLAST against each other. It was found that over the complete dataset only 5281 did match with a total of 3995 contigs detected as known proteins. The program BLAST2GO was used to map the top BLASTx matches (*E*-value 10^{-6}) and to assign gene ontology (GO) term annotations. Overall, there was little difference in the GO assignments for the three assemblies (Table 2). In both systems, the distribution of the contigs in various functional classes of the GO database indicates broad gene diversity.

Acknowledgement

The work was supported by the International Foundation for Science (IFS Grant Agreement No.: A/4914/1), the National Research Foundation (NRF Grant No.: 71069) of South Africa and by the Claude Leon Foundation (Grant No.: 1407). The authors have no conflict of interests.

References

[1] V.D.R. Falcao, A.P. Tonon, M.C. Oliveira, P. Colepicolo, RNA isolation method for polysaccharide rich algae: agar producing *Gracilaria tenuistipitata* (Rhodophyta). J. Appl. Phycol. 20 (1) (2008) 9–12.

[2] P. Franchini, M. van der Merwe, R. Roodt-Wilding, Transcriptome characterization of the South African abalone *Haliotis midae* using sequencing-by-synthesis. BMC Res. Note 4 (2011) 59.

[3] E.A. Hornett, C.W. Wheat, Quantitative RNA-seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. BMC Genomics 13 (1) (2012) 361–376.

[4] B. Picone, C. Rhode, R. Roodt-Wilding, Transcriptome profiles of wild and cultured South African abalone, *Haliotis midae*. Mar. Genomics 20 (2015) 3–6.

[5] M. Van der Merwe, Growth-Related Gene Expression in *Haliotis midae*. Stellenbosch University, South Africa, 2010 (PhD dissertation).

[6] Y. Wang, N. Ghaffari, C.D. Johnson, U.M. Braga-Neto, H. Wang, R. Chen, H. Zhou, Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinform. 12 (10) (2011) 5–10.