manuscript

Yuanning Li

10/17/2018

# Lamellibrachia genome

## Abstract

## Introduction

 **Fig 1.** *Lamellibrachia* from seep localities in Gulf of Mexico.

## Methods

### Biological materials.

Adult *Lamellibrachia* *luymesi* specimens were collected from from seep localities in the Mississippi Canyon at 754 m depth in Gulf of Mexico (N 28°11.58’, W 89°47.94’), using the R/V Seward Johnson and Johnson Sea Link in October 2009. All samples were frozen at 80˚C following collection.

### Genome sequencing and assembly.

Vestimentum tissue was dissected from one individual of worm, and high molecular weight genomic DNA was extracted using the the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer’s protocols. Sequencing a total of six paired-end or mate-pair genomic DNA libraries with insert sizes ranging from 180 bp to 7 kb were performed by by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama on an Illumina HiSeq 2000 platform (see details in Table S1). Paired-end libraries (180 bp, 400 bp, 750 bp) were prepared using the 125 bp TrueSeq protocols, and mate-pair libraries (3-5 kbp, 5-7 kbp) were generated using the Illuomina Nextera Mate Pair Library Kit followed by size selection. Moreover, a 10X sequencing library was constructed using the 10X Chromium protocol (10X genomics) at the Hudson Alpha Institute. The finished library were sequenced on an Illumina HiSeqX platform, using paried 151 bp reads with a single 8 bp index read.

Our workflow of genome assembly was shown in Fig. S1. The paired-end and 10X raw reads were checked using FastQC v0.11.5 (Andrews and others 2010) and quality filtered (Q score >30) using Trimmomatic v0.36 (Bolger, Lohse, and Usadel 2014). The estimatation of genome size, level of heterozygosity and repeat contes of the Lamellibrachia genome was determined by analaysing the kmer histograms generated from the paired-end librries using Jellyfish v2.2.3 (Marçais and Kingsford 2011) and GenomeScope (Vurture et al. 2017) (Fig. S2). The Mate-pair reads were trimmed and sorted using NxTrim v0.3.1 (O’Connell et al. 2015) which can recgonize and trim the artificial Nextera mate-pair circulation adapters. Only reads from category “mp” (true mate-pair reads) and “unkonwn” (mostly large insert size reads) were used for downstream scaffolding anlaysis. Reads from “pe” (paired-end reads) and “se” (single ends) categories were discarded.

Given that high heterozygosity of *Lamellibrachia* genome, all reads were assembled using Platanus v1.2.4 (Kajitani et al. 2014) with a kmer size of 32. Scaffolding was conducted by mapping Illumina paired-end and mate-pair reads to contigs genrated by Platanus using SSPACE v3.0 (Boetzer and Pirovano 2014). Gaps in the scaffolds were then filled with GapCloser v1.12 (Luo et al. 2012). Redundant allele scaffods were further remvoed using Redundans v0.13c with default settings (Pryszcz and Gabaldón 2016). Genome assembly quality was assessed using QUAST v3.1 (Gurevich et al. 2013). Completeness of obtained genome was assessed using BUSCO v3(Waterhouse et al. 2017) with Metazoa\_odb9 database (978 busco genes) .

### Genome annotation.

Our genome annotation workflow was shown in Fig. S3. Gene models of *Lamellibrachia* genome were constructed following the Funannotate pipeline 1.3.0 (<https://github.com/nextgenusfs/funannotate>). Briefly, repeptive regions in the *Lamellibrachia* genome were identified usning RepeatModeler v1.0.8 (Smit and Hubley 2008) and were subsequently soft-masked using RepeatMasker v4.0.6 (Chen 2004). RNA-Seq data from different tissue were leveraged to improve the accuracy of gene prediction. RNA-Seq data were assembled *de* *novo* into transcriptomes using Trinity v2.4.0 (Haas et al. 2013) and HISAT 2.1.0 (Kim, Langmead, and Salzberg 2015) was used to algin RNA-Seq reads to the *Lamellibrachia* assembly. Transcrptome assemblies were then passed to PASA pipeline v2.3.3 (Haas et al. 2003) to identify high quality gene models. The aligned RNA-Seq data wes used to train the *ab* *initio* gene predictions using AUGUSTUS v3.3 (Stanke et al. 2006). Protein alignements from the SwissProt database to “Lamellibrachia” assembly were generated using exonerate (Slater and Birney 2005) and Trinity/PASA transcripts were aligned to the genome using Minimap2 v2.1 (Li 2018). The tRNA genes were identified using tRNAscan-SE v1.3.1 (Lowe and Eddy 1997). EvidenceModeler 1.1.0 (Haas et al. 2008) was used to combine all the evidences of gene prediction from protein alignemnts, transcritp alignments, and *ab* *initio* predictions to construct high quality gene models. Finally, functional annotations of predicted gene models were analyzed using several curated databases. KEGG orthology was assinged using the KEGG Automatic Annotation server. Gene models were further annotated with domain structure and protein identity by InterProScan (Zdobnov and Apweiler 2001)and SwissProt database, respectively. Secreted proteins were predicted using SignalP (Petersen et al. 2011) and Phobius (Käll, Krogh, and Sonnhammer 2007) using InterProScan.

## Results and Discussion

## Conclusion

## Ackowledgements

## Author contribution

## Supplemental Information

Andrews, Simon, and others. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.”

Boetzer, Marten, and Walter Pirovano. 2014. “SSPACE-Longread: Scaffolding Bacterial Draft Genomes Using Long Read Sequence Information.” *BMC Bioinformatics* 15 (1): 211.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.

Chen, Nansheng. 2004. “Using Repeatmasker to Identify Repetitive Elements in Genomic Sequences.” *Current Protocols in Bioinformatics* 5 (1): 4–10.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–5.

Haas, Brian J, Arthur L Delcher, Stephen M Mount, Jennifer R Wortman, Smith JrRoger K, Linda I Hannick, Rama Maiti, et al. 2003. “Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies.” *Nucleic Acids Research* 31 (19): 5654–66.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from Rna-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494.

Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. 2008. “Automated Eukaryotic Gene Structure Annotation Using Evidencemodeler and the Program to Assemble Spliced Alignments.” *Genome Biology* 9 (1): 1.

Kajitani, Rei, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, et al. 2014. “Efficient de Novo Assembly of Highly Heterozygous Genomes from Whole-Genome Shotgun Short Reads.” *Genome Research*, gr–170720.

Käll, Lukas, Anders Krogh, and Erik LL Sonnhammer. 2007. “Advantages of Combined Transmembrane Topology and Signal Peptide Prediction—the Phobius Web Server.” *Nucleic Acids Research* 35 (suppl\_2): W429–W432.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg. 2015. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” *Nature Methods* 12 (4): 357.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 1: 7.

Lowe, Todd M, and Sean R Eddy. 1997. “TRNAscan-Se: A Program for Improved Detection of Transfer Rna Genes in Genomic Sequence.” *Nucleic Acids Research* 25 (5): 955.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. “SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler.” *Gigascience* 1 (1): 18.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27 (6): 764–70.

O’Connell, Jared, Ole Schulz-Trieglaff, Emma Carlson, Matthew M Hims, Niall A Gormley, and Anthony J Cox. 2015. “NxTrim: Optimized Trimming of Illumina Mate Pair Reads.” *Bioinformatics* 31 (12): 2035–7.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8 (10): 785.

Pryszcz, Leszek P, and Toni Gabaldón. 2016. “Redundans: An Assembly Pipeline for Highly Heterozygous Genomes.” *Nucleic Acids Research* 44 (12): e113–e113.

Slater, Guy St C, and Ewan Birney. 2005. “Automated Generation of Heuristics for Biological Sequence Comparison.” *BMC Bioinformatics* 6 (1): 31.

Smit, AFA, and R Hubley. 2008. “RepeatModeler Open-1.0.” *Available Fom Http://Www. Repeatmasker. Org*.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts.” *Nucleic Acids Research* 34 (suppl\_2): W435–W439.

Vurture, Gregory W, Fritz J Sedlazeck, Maria Nattestad, Charles J Underwood, Han Fang, James Gurtowski, and Michael C Schatz. 2017. “GenomeScope: Fast Reference-Free Genome Profiling from Short Reads.” *Bioinformatics* 33 (14): 2202–4.

Waterhouse, Robert M, Mathieu Seppey, Felipe A Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V Kriventseva, and Evgeny M Zdobnov. 2017. “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.” *Molecular Biology and Evolution* 35 (3): 543–48.

Zdobnov, Evgeni M, and Rolf Apweiler. 2001. “InterProScan–an Integration Platform for the Signature-Recognition Methods in Interpro.” *Bioinformatics* 17 (9): 847–48.