manuscript

Yuanning Li

10/17/2018

# Lamellibrachia genome

## Abstract

## Introduction

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

## Methods

Scripts and data for the analyses are available in a git repository at <https://github.com/yzl0084/Lamellibrachia-genome>.

### 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. Two 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).

### Transcripome assembly and analysis

Total RNA was extracted from the plume, vestimentum and trophosome tissue from the same indivdisual of *Lamellibrachia* spicemen using Trizol. RNA-seq of adult tissues from plume, vestimentum and trophosome was performed using Illumina HiSeq 2000 platform in Hudson Alpha. After quality checking and trimming of raw sequencing reads, transcripts were assembled de novo with Trinity. Transcript isoforms with high similarity (≥ 95%) were removed with CD-HIT-EST v4.7 (Li and Godzik 2006). Transcript abundance was estimated with Bowtie v2.2.9 (Langmead and Salzberg 2012) and RSEM v1.2.26 (Li and Dewey 2011) by mapping reads back to the transcriptomic assembly based on transcripts per million (TPM). A tissue specifically expressed gene was defined as a gene that had over 75% of the total transcripts in a particular tissue based on TPM (Albertin et al. 2015). GO enrichment analysis was performed based on the GO annotation using PANNZER2 (Törönen, Medlar, and Holm 2018). Statistically overrepresented GO terms of trophosome-specific genes were identifed using Lamellibrachia gene models as background with AgriGO (Du et al. 2010).

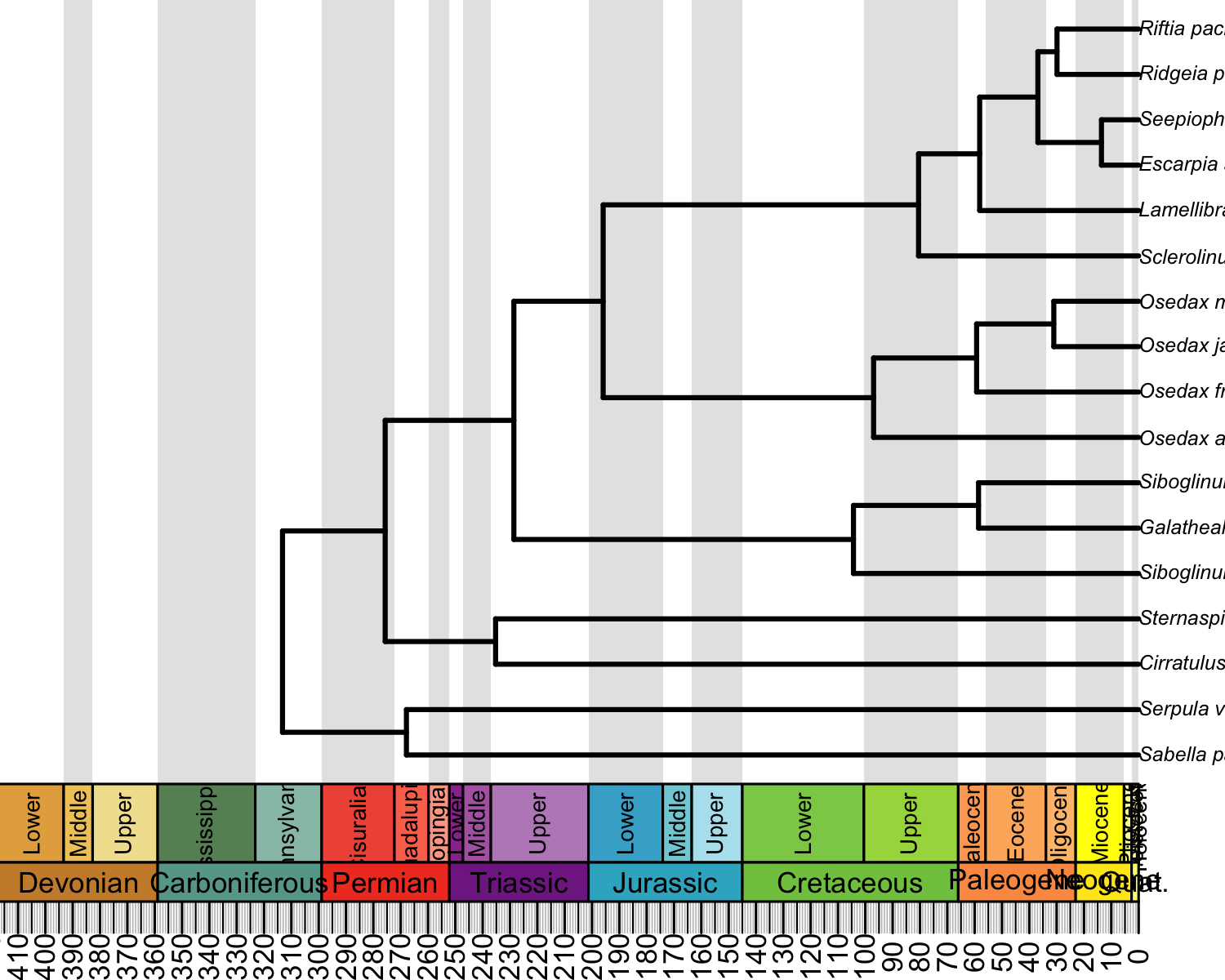
### Genome annotation.

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). Finally, 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.

### Phylogenomics adn Molecular clock analysis

Analysis of siboglinid phylogeny was conducted from publically available 16 transcriptomic data in conjuction with *Lamellibrachia* proteome. Sequence assembly, annotation, homology evaluation, gene tree construction, parsing of genes trees to OGs, and supermatrix construction were conducted with Agalma (Dunn, Howison, and Zapata 2013). The final analyses presented here contained 13 siboglinids and four outgroups based on current understanding of annelid phylogeny. Maximum likelihood analyses were performed in IQTree under the best-fitting models for associated partition schemes determined by Modelfinder implemented in IQTree 1.6.3 (Nguyen et al. 2014) with ultrafast bootstrapping of 1000 replicates.

For molecular clock analysis, a relaxed molecular clock with a lognormal distribution and a Yule tree model was used in BEAST 2 v2.5.1 (Bouckaert et al. 2014). A calibration was placed on the node representing the most recent common ancestor (MRCA) of *Osedax* using a normal distribution with a mean of 100 MYA and a standard deviation of 10 following the findings of (Danise and Higgs 2015).Another calibrartion was placed on the node of MRCA of Serpulida and Sabellida using a normal distribution with a mean of 267 MYA (Sanfilippo et al. 2017). Molecular clock analyses with BEAST 2 consisted of two independent runs with 1 million MCMC generations sampled every 1000 generations. Convergence was checked and confirmed by comparing trace plots in Tracer making sure the effective sample size of each parameter was greater than 100 and that stationarity appeared to have been achieved. Log and tree fiels were combined using Logcombiner. A maximum clade credibility tree with mean heights was calculated using TreeAnnotater. The resulted time-calibrated tree was plotted using R package, phyloch, strap and OutbreakTools. Bayesian inference using a molecular clock resulted in identical branching patterns as analysis with IQTree.



### Gene family analysis

After all-to-all Diamond v1.0 (Buchfink, Xie, and Huson 2014) BLASTP searches against 12 selecte lophotrochozoan proteomes, orthology groups (OGs) were identified using Orthofinder with a dfault inflation parameter (I=1.5). Gene ontology annotation was performed using PANTHER v13.1 (Mi et al. 2016) with the PANTHER HMM scoring tool (pantherScore2.pl). Gene family expansion and contraction was estimated using CAFÉ v2.1 (De Bie et al. 2006). For each gene family, CAFÉ generated a family-wide P value, with a significant *P* value indicating a possible gene-family expansion or contraction event. A branch-specific *P* value was also generated for each branch/node using the Viterbi method. A family-wide *P* value less than 0.01 and a branch/node Viterbi *P* value less than 0.001 was considered as a signature of gene family expansion/contraction for a specific gene family and specific species, respectively, as suggested in manual.

### Hemoglobin and immunity-related genes.

## Results and Discussion

### Genome features

Results from high-throughput sequencing and genome assembly for *Lamellibrachia* are presented in Table # with at least 300 fold coverage using a combination of Illumina paired-end, mate-pair and 10X genomic sequencing with a variety of insert sizes. The haploid genome assembly size is ~687 Mb, similar to the result of genome size estimation using GenomeScope (~646 Mb). The N50 value of the assembled scaffolds and contigs is 373 Kb and 24 Kb, respectively. Although N50 lengths and assembly quality of *Lamellibrachia* are comparable to those of other annelids (e.g. *Capitella* *teleta*, *Hebdella* *robusta*), the overall genome completeness measured by BUSCO (~ 95%) is one of the highest among other lophotrochozoans (Table #), indicating the completeness of the genome assembly. With the support of RNA-seq data from different tissue, we estimated *Lamellibrachia* genome contains 38,998 gene models. The genome also exhibit similar level of heterozygosity (0.6%) and repeptitive seuqences (36.92%), compared to other lophotrochzoan genomes (Table #). Comparison among three annelid genome revealed

### Phylogeny and molecular clock of tubeworm

Two additional *Osedax* taxa were added in the study compared to the previous siboglinid phylogenomic analyses (Li et al. 2017) (Table #). The final supermatrix dataset contains 191 genes single-copy orthologs. Bayesian inference with a relaxed molecular clock (Fig. 2) recovered the same topology as ML analysis in IQTree with strong nodal support (Fig. #). Both analyses strongly support *Osedax* is most closely related to the Vestimentifera + *Sclerolinum* clade and Frenulata is the early diverging group, as recently reported (Li et al. 2015). Within Vestimentifera, *Lamellibrachia* is sister to the remaining sampled vestimentiferans.

Previous Molecular data indicate a Late Mesozoic or Cenozoic (approx. 50–100 Myr) origin for the siboglinids [1,22] based on COI

## Conclusion

## Ackowledgements

## Author contribution

## Supplemental Information

Albertin, Caroline B, Oleg Simakov, Therese Mitros, Z Yan Wang, Judit R Pungor, Eric Edsinger-Gonzales, Sydney Brenner, Clifton W Ragsdale, and Daniel S Rokhsar. 2015. “The Octopus Genome and the Evolution of Cephalopod Neural and Morphological Novelties.” *Nature* 524 (7564): 220.

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.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. 2014. “BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.” *PLoS Computational Biology* 10 (4): e1003537.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. “Fast and Sensitive Protein Alignment Using Diamond.” *Nature Methods* 12 (1): 59.

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

Danise, Silvia, and Nicholas D Higgs. 2015. “Bone-Eating Osedax Worms Lived on Mesozoic Marine Reptile Deadfalls.” *Biology Letters* 11 (4): 20150072.

De Bie, Tijl, Nello Cristianini, Jeffery P Demuth, and Matthew W Hahn. 2006. “CAFE: A Computational Tool for the Study of Gene Family Evolution.” *Bioinformatics* 22 (10): 1269–71.

Du, Zhou, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. 2010. “AgriGO: A Go Analysis Toolkit for the Agricultural Community.” *Nucleic Acids Research* 38 (suppl\_2): W64–W70.

Dunn, Casey W, Mark Howison, and Felipe Zapata. 2013. “Agalma: An Automated Phylogenomics Workflow.” *BMC Bioinformatics* 14 (1): 330.

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.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357.

Li, Bo, and Colin N Dewey. 2011. “RSEM: Accurate Transcript Quantification from Rna-Seq Data with or Without a Reference Genome.” *BMC Bioinformatics* 12 (1): 323.

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

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–9.

Li, Yuanning, Kevin M Kocot, Christoffer Schander, Scott R Santos, Daniel J Thornhill, and Kenneth M Halanych. 2015. “Mitogenomics Reveals Phylogeny and Repeated Motifs in Control Regions of the Deep-Sea Family Siboglinidae (Annelida).” *Molecular Phylogenetics and Evolution* 85: 221–29.

Li, Yuanning, Kevin M Kocot, Nathan V Whelan, Scott R Santos, Damien S Waits, Daniel J Thornhill, and Kenneth M Halanych. 2017. “Phylogenomics of Tubeworms (Siboglinidae, Annelida) and Comparative Performance of Different Reconstruction Methods.” *Zoologica Scripta* 46 (2): 200–213.

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.

Mi, Huaiyu, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D Thomas. 2016. “PANTHER Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements.” *Nucleic Acids Research* 45 (D1): D183–D189.

Nguyen, Lam-Tung, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2014. “IQ-Tree: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

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.

Sanfilippo, Rossana, Antonietta Rosso, Agatino Reitano, and Gianni Insacco. 2017. “First Record of Sabellid and Serpulid Polychaetes from the Permian of Sicily.” *Acta Palaeontologica Polonica* 62 (1): 25–38.

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.

Törönen, Petri, Alan Medlar, and Liisa Holm. 2018. “PANNZER2: A Rapid Functional Annotation Web Server.” *Nucleic Acids Research*.

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.