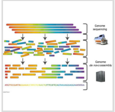

Apr 21, 2019

Working

The pipeline of assembly and annotation of the Scapharca broughtonii genome

In 1 collection

Chang-Ming Bai[1]

[1]Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences

dx.doi.org/10.17504/protocols.io.z7zf9p6

Chang-Ming Bai

ABSTRACT

This protocol include the detailed methods of assembly and annotation of the *Scapharca broughtonii* genome.

1 The original data generated with PacBio and Nanopore platforms were base-called, quality controlled, combined and transferred to FASTA format for assembly.

2 Run Canu (v1.5) for long reads correction, triming and assembly.

correctedErrorRate=0.045, corOutCoverage=6.

3 Run Wtdbg (v1.1) for further assembly of the data obtained in step 2.

-k 21(Kmer size:21), -e 3(Min cov of edges=3q), the other parameter was set as default.

4

Run Quickmerge (v0.2.2) to combine assembled results of step 2 and 3, then run Numer (v4.0.0) to remove the redundancy, and finally run Pilon (v1.22) to correct sequencing errors in the assembly with Illumina reads.

Quickmerge: -hco 5.0 (controls the overlap cutoff used in selection of anchor contigs. Default is 5.0) -c 1.5 controls the overlap cutoff for contigs used for extension of the anchor contig. Default is 1.5) -l 400000 (controls the length cutoff for anchor contigs) -ml 5000 (controls the minimum alignment length to be considered for merging);
Numer: default parameters;
Pilon: -mindepth 10, the other parameter was set as default.

5 "Run SAMTools (v0.1.18) to evaluate the assembly quality by mapping the 360,937,442 Illumina reads for genome survey to the assembly.

Run BUSCO (v2.0) to evaluate the assembly quality by searching the 303 eukaryotic and 978 metazoan conserved genes in the assembly."

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

SAMTools: no parameter;
BUSCO: default parameters.

6 Run LTR FINDER (v1.05), RepeatScout (v1.0.5) and PILER-DF (v2.4) to build a de novo library.

All with default parameters.

7 Run PASTEClassifier (V1.0) to classify the repetitive sequences in the library constructed in step 5, and cnmbined with Repbase database to creat the final library.

Default parameters

8 Basing on the library constructed in step 6 as database, run RepeatMasker (v4.0.6) to identify repeats in the genome.

-nolow -no_is -norna -engine wublast -qq -frag 20000

9 Mask these repetitive regions obtained above (step 5-7) with 'N's.

Preparation for gene prediction.

10 "Download transcriptome data of the S. broughtonii uploaded by us from NCBI. Illumina data was assembled by Trinity (v.r20140413p1) in previous study, and reassembled by Hisat (v2.0.4) and Stringtie (v1.2.3);
Pacbio data was full-length transcripts obtained after quality control;
Run TransDecoder v2.0 (http://transdecoder.github.io) and GeneMark (v5.1) to predict the gene functions.

Trinity: min_kmer_cov:2, and set the other parameters as default;
Hisat: default parameters;
Stringtie: default parameters;
TransDecoder: default parameters;
GeneMark: default parameters.

11 Run GeMoMa (v1.3.1) for homology-based prediction by aligning the assembled genome against those of 4 closely related species (Danio rerio, Crassostrea gigas, Mizuhopecten yessoensis and Mytilus galloprovincialis) downloaded from NCBI.

Default parameters

12 Run Augustus (v. 2.4), Genscan (v. 3.1), GlimmerHMM (v3.0.4), GeneID (v1.4) and SNAP (version 2006-07-28) to de novo predict genes in the repeat-masked genome sequence.

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

All with default parameters

13 Run EVM (v1.1.1) to obtain the consensus gene set by integrating genes predicted in step 10-12.

STANDARD S-ratio: 1.13 score>1000
Weights used for predicted genes by different softers are list below:
PROTEIN OTHER 50
PROTEIN GeMoMa 50
TRANSCRIPT assembler-PASA 50
TRANSCRIPT Stringtie 20
ABINITIO_PREDICTION genscan 0.3
ABINITIO_PREDICTION AUGUSTUS 0.3
ABINITIO_PREDICTION GlimmerHMM 0.3
ABINITIO_PREDICTION SNAP 0.3
ABINITIO_PREDICTION geneID 0.3
ABINITIO_PREDICTION GeMoMa 0.3
OTHER_PREDICTION OTHER 100

14 Run PASA v2.0.2 to modify the genes predicted in step 13.

Default parameters

15 "Map protein sequences of the final gene set to existing databases to identify their functions or motifs, such as Nr, Nt, SwissProt, TrEMBL, KOG, KEGG, Pfam and GO.
Nr and Nt were downed on 2017.04.05, the other databases were downed on 2017.02.13.

Nr, Nt, SwissProt, TrEMBL, KOG and KEGG: using BLAST (v2.2.31);
Blast: -max_target_seqs 100
Pfam: using HMMer (v3.0), -E 0.00001 --domE 0.00001, , the other parameter was set as default;
GO: using BLAST2GO (v2.5) with default parameters.

16 Run genBlastA (v1.0.4) to search for putative pseudogenes based on homology, and run GeneWise (v2.4.1) to identify pseudogenes based mutations.

genBlastA: -e 1e-5
GeneWise: -both -pseudo

17 Run tRNAscan-SE (v1.3.1) to predict tRNA.

Default parameters

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

18 Run Infernal (v1.1) to predict miRNA and rRNA based on Rfam v12.1 and miRBase v21.0 databases.

Default parameters

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited