Whole genome de-novo assembly and annotation protocol for Apostichopus japonicus genome

Jihoon Jo, Jooseong Oh, Hyun Gwan Lee, Hyun Hee Hong, Sung Gwon Lee, Seongmin Cheon, Elizabeth MA Kern, Soyeong Jin, Sung Jin Cho, Joong Ki Park, Chungoo Park

Abstract

This protocol is for the whole genome de novo assembly and annotation for apostichopus japonicus genome, but can be useful in other marine invertebrate genomes. It accompanies the following *GigaScience* publication:

Jihoon Jo, et al. (2016): Draft genome of the sea cucumber Apostichopus japonicus and genetic polymorphism among color variants. *GigaScience*...

Citation: Jihoon Jo, Jooseong Oh, Hyun Gwan Lee, Hyun Hee Hong, Sung Gwon Lee, Seongmin Cheon, Elizabeth MA Kern, Soyeong Jin, Sung Jin Cho, Joong Ki Park, Chungoo Park Whole genome de-novo assembly and annotation protocol for Apostichopus japonicus genome. **protocols.io**

dx.doi.org/10.17504/protocols.io.gmabu2e

Published: 01 Dec 2016

Protocol

Step 1.
Run Trimmomatic and "Trim Galore" with input files of all Illumina raw-reads.

Step 2.
Run ALLPATHS-LG for correcing errors in the raw illumina sequences

Step 3.
Run Platanus for de-novo genome assembly

NOTES
GigaScience Database 30 Nov 2016

You should compare the performance of assemblers such as SOAPdenovo, ALLPATHS-LG

Step 4.
Run GapCloser (a module of SOAPdenovo2) for gap closing of the gaps that still remained in the resulting scaffolds

Step 5.
Run CEGMA

Step 6.
Run BUSCO

Step 7.

Run RepeatMasker using the Repbase TE library and de novo repeat library constructed by Repeat Modeler

Step 8.
Perform the ab-initio gene prediction using AUGUSTUS using hints from splicing alignment of transcripts to the repeat-masked assembled genome with BLAT and PASA.

NOTES
GigaScience Database 30 Nov 2016

You must collect the RNA-seq data for obtaing AUGUSTUS training set

Gene prediction

Step 9.

Map the homologous proteins in other species (from Unoprot) to the repeat-masked assembled genome using tBLASTn with an E-value < 1×10^{-5}

Step 10.
Predict the gene in the aligned sequences in Step9 using GeneWise to search for precise spliced alignment and gene structure.

Step 11.
Map RNA-seq reads to the repeat-masked assembled genome using TopHat, and build the homology-based gene model using Cufflinks.

Step 12.
Intergrate all gene set in Step 8-11 without redundant gene set.