Aug 23, 2018

Fish genome assembly and annotation pipeline

GigaScience

In 1 collection

DOI
dx.doi.org/10.17504/protocols.io.ss3eegn

Chang Li[1]

[1]BGI-Qingdao, BGI-Shenzhen, Qingdao, 266555, China.

GigaScience Press BGI

Chang Li

DOI: dx.doi.org/10.17504/protocols.io.ss3eegn

External link: https://doi.org/10.1093/gigascience/giy114

Protocol Citation: Chang Li 2018. Fish genome assembly and annotation pipeline. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.ss3eegn**

Manuscript citation:
hangwei Shao, Chang Li, Na Wang, Yating Qin, Wenteng Xu, Qun Liu, Qian Zhou, Yong Zhao, Xihong Li, Shanshan Liu, Xiaowu Chen, Shahid Mahboob, Xin Liu, Songlin Chen, Chromosome-level genome assembly of the spotted sea bass, *Lateolabrax maculatus*, *GigaScience*, Volume 7, Issue 11, November 2018, giy114, https://doi.org/10.1093/gigascience/giy114

Protocol status: Working
We use this protocol and it's working

Created: August 22, 2018

Last Modified: August 24, 2018

Protocol Integer ID: 14907

Abstract

From this protocol, we can know detail methods of assembly and annotation of the L. maculatus genome.

Quality control

1 Get raw sequencing data in Fastq format. Filter the input raw sequences by using SOAPnuke (v.1.5.6).

Note

using parameters " -l 5 -q 0.5 -n 0.1 -Q 2 --seqType 0"

k-mer analysis

2 Estimate the genome size (650 Mb) with k-mer analysis.

Note

k=17; about 27.7Gb reads from as input; the genome size with the formula: G = N*(L − 17 + 1)/K_depth, where N and L are the total number of reads and the length of reads, respectively, and K_depth indicates the frequency of k-mers occurring more frequently than the others.

Assembly

3 1) Run SOAPdenovo2(v. 2.04.4) to assemble our genome.
2) Perform krskgf (v. 1.19) and Gapcloser (v. 1.10) to further close gaps in our genome obtained in step3.

Note

1) performing "pregraph (-K 57 -p 10)->contig (-g)->map (-p 10 -k 39)->scaff(-p 10)" modes in turn;
2) using reads from all insert-size libraries.

Repeat annotation_de novo

4 1) Run RepeatModeler to build de novo library based on the input assembled genome sequence.
2) Basing on the library constructed in step 5 as database, run RepeatMasker (v. 3.3.0) to find and then classify the repetitive sequences.

Note

2) using parameters "-nolow -no_is -norna -parallel 1"

Repeat annotation_database

5 Run TRF (v. 4.09), RepeatMasker and RepeatProteinMask (v. 3.3.0) to identify repeats in the genome at DNA and protein level, respectively, by aligning sequences against Repbase library (v. 17.01).

Note

using parameters "-noLowSimple -pvalue 0.0001" when running RepeatProteinMask

Gene prediction_preparation

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

Note

Before gene prediction, mask the TEs in genome.

Gene prediction_de novo

7 Run Augustus (v. 2.5.5) and Genscan (v. 2.1) to de novo predict genes in the repeat-masked genome sequences.

Note

using parameters "--species=Lateolabrax_maculatus --uniqueGeneId=true --noInFrameStop=true --gff3=on --strand=both" when running Augustus; using default parameters when runing Genescan.

Gene prediction_homolog

8 Download protein sequences of teleost species (Danio rerio (NCBI, GenBank ID:50), D. labrax (NCBI, GenBank ID:2659), Gasterosteus aculeatus (NCBI, GenBank ID:146), Lates calcarifer (NCBI, GenBank ID:14180), Oreochromis niloticus (NCBI, GenBank ID:197), Oryzias latipes (NCBI, GenBank ID:542), Tetraodon nigroviridis (NCBI, GenBank ID:191) and Takifugu rubripes (NCBI, GenBank ID:63)), then align these against our masked

genome sequences with BLAT, and then based on the BLAT mapping results, run GeneWise (v. 2.2.0) to predict genes.

Note

with parameters "--max divergence rate 0.3 --extend length for both sides of regions 2000"

Gene prediction_transcriptome

9 Download transcriptome data of the spotted sea bass from NCBI. The data was assembled by Trinity (v.2013.11). Mapp this sequence to our genome with BLAT and then based on the BLAT results, reduced redundancy genes.

Note

default parameters

Gene prediction_GLEAN

10 Integrate genes predicted in step 9-11 to obtain the consensus gene set by using GLEAN.

Note

filtering with criterion "overlap cutoff 0.8 and at least one homolog support"

Final gene set

11 Added the genes which were supported by the transcriptome data and D. labrax's based prediction after manual evolution to the GLEAN gene set.

Functional annotation

12 Map protein sequences of the final gene set to existing databases to identify their functions or motifs, such as SwissProt, TrEMBL, KEGG, InterPro.

Note

SwissProt, TrEMBL and KEGG: using BLASTP; Interpro: using InterProScan (version 4.7) with seven different models (Profilescan, blastprodom, HmmSmart, HmmPanther, HmmPfam, FPrintScan and Pattern-Scan)