The pipeline of assembly and annotation

Yuanyuan Fu, Liangwei Li, Shijie Hao, Rui Guan, Guangyi Fan, Chengcheng Shi, Haibo Wan, Wenbin Chen, He Zhang, Guocheng Liu, Jihua Wang, Lulin Ma, Jianling You, Xuemei Ni, Zhen Yue, Xun Xu, Xiao Sun, Xin Liu, Simon Ming-Yuen Lee

Abstract

This protocol provides the detailed methods of assembly and annotation of the R. crenulata genome.

Citation: Yuanyuan Fu, Liangwei Li, Shijie Hao, Rui Guan, Guangyi Fan, Chengcheng Shi, Haibo Wan, Wenbin Chen, He Zhang, Guocheng Liu, Jihua Wang, Lulin Ma, Jianling You, Xuemei Ni, Zhen Yue, Xun Xu, Xiao Sun, Xin Liu, Simon Ming-Yuen Lee The pipeline of assembly and annotation. **protocols.io**

https://www.protocols.io/view/the-pipeline-of-assembly-and-annotation-hrpb55n

Published: 19 May 2017

Before start

Get raw sequencing data in Fastq format.

Protocol

Quality contol

Step 1.

Filter the input raw sequences by using SOAPfilter (version 2.2).

SOFTWARE PACKAGE (linux)

SOAP: short oligonucleotide alignment program, 2.2

BGI

http://soap.genomics.org.cn/#down2

NOTES

GigaScience Database 25 Apr 2017

using parameters "-y -z -p -M 2"

k-mer analysis

Step 2.

Estimate the genome size (420 Mb) with k-mer analysis.

NOTES

GigaScience Database 25 Apr 2017

k=17; reads from 250 bp-insert library 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

Step 3.

Run Platanus (version 1.2.4) to assemble our genome.

SOFTWARE PACKAGE (linux)

Platanus Assembler (PLATform for Assembling NUcleotide Sequences, 1.2.4

Tokyo Institute of Technology
http://platanus.bio.titech.ac.jp/platanus/?page_id=14

cmd COMMAND

```
assemble (-k 35 -m 130 -u 0.2 -c 10 -s 10 -t 32)

scaffold (-l 3 -u 0.2 -v 32 -s 32)

gap_close (-t 32 -s 32 -ed 0.1)
```

(Perform assemble -> scaffold -> gap_close modes in turn)

NOTES

GigaScience Database 25 Apr 2017

performing "assemble (-k 35 -m 130 -u 0.2 -c 10 -s 10 -t 32)->scaffold (-l 3 -u 0.2 -v 32 -s 32)->gap_close (-t 32 -s 32 -ed 0.1)" modes in turn;

Assembly

Step 4.

Perform Gapcloser (version 1.10) to further close gaps in our genome obtained in step3.

SOFTWARE PACKAGE (linux)

GapCloser, 1.10

BGI
http://soap.genomics.org.cn/soapdenovo.html

NOTES

GigaScience Database 25 Apr 2017

using reads from all insert-size libraries

Repeat annotation_de novo

Step 5.

Run RepeatModeler, RepeatScout and LTR_FINDER, respectively, to build de novo library based on the input assembled genome sequence.

NOTES

GigaScience Database 25 Apr 2017

version 1.0.5

Repeat annotation_de novo

Step 6.

Based on the library constructed above as database, run RepeatMasker (version 3.3.0) to find and then classify the repetitive sequences.

SOFTWARE PACKAGE (linux)

RepeatMasker, 3.3.0

Institute for Systems Biology

NOTES

GigaScience Database 25 Apr 2017

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

Repeat annotation_homolog

Step 7.

Run RepeatMasker and RepeatProteinMask (version 3.3.0) to identify repeats in the genome at DNA and protein level, respectively, by aligning sequences against existing databases, Repbase TE library (Version 17.01) and TE protein database.

NOTES

GigaScience Database 25 Apr 2017

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

Gene prediction_preparation

Step 8.

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

Gene prediction_de novo

Step 9.

Run Augustus (version 2.5.5) and GlimmerHMM (version 3.0.1) to de novo predict genes in the repeat-masked genome sequences.

NOTES

GigaScience Database 25 Apr 2017

using parameters "--species=arabidopsis --uniqueGeneId=true --noInFrameStop=true --gff3=on --strand=both" when running Augustus; using parameters "-f -g" and trained by arabidopsis when running GlimmerHMM

Gene prediction_homolog

Step 10.

Download protein sequences of homlog species (arabidopsis, wild strawberry, peach and Chinese plum), then align these against our masked genome sequences with BLAT, and then based on the

BLAT mapping results, run GeneWise (version 2.2.0) to predict genes.

NOTES

GigaScience Database 25 Apr 2017

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

Gene prediction_glean

Step 11.

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

NOTES

GigaScience Database 25 Apr 2017

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

Gene prediction_adding

Step 12.

Perform TopHat (version 2.1.0) with default parameters to align clean RNA-seq reads against gene set mentioned in Step11, and then use Cufflinks (version 2.2.1) to assemble these transcripts, then use training parameters to predict ORFs, and finally obtain the more intergrity and trusty gene set.

NOTES

GigaScience Database 25 Apr 2017

filtering RNA sequencing data firstly by performing SOAPnuke with parameters “-l 10 -q 0.5 -n 0.01 -f AGTCGGAGGCCAAGCGGTCTTAGGAAGACAA -Q 2”

Estimation of completeness

Step 13.

Run BUSCO and map transcriptome data to our final gene set to assess the completeness.

Functional annotation

Step 14.

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

NOTES

GigaScience Database 25 Apr 2017

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)