An analytical pipeline of assembly and annotation of the Betta splendens genome.

Xin Liu

Abstract

From here, You can learn about the detail methods of genome assembly and gene annotation of the Betta splendens genome.

Citation: Xin Liu An analytical pipeline of assembly and annotation of the Betta splendens genome.. **protocols.io** dx.doi.org/10.17504/protocols.io.qq9dvz6
Published: 08 Jun 2018

Protocol

Quality Control

Step 1.

Get raw sequencing data in Fastq format. Filter the raw sequencing data by using SOAPfilter (version 2.2).

NOTES
Hongling Zhou 06 Jun 2018

Using parameters '-i insertsize -y -z -p -M 2'

k-mer analysis

Step 2.

Estimate the genome size with k-mer (version 1.0) analysis.

NOTES
Hongling Zhou 06 Jun 2018

Using parameters 'k=17 -t 12'

Assembly

Step 3.

1. Run SOAPdenovo (version 2.04) to assemble the Betta splendens genome.

2. Perform Gapcloser (version 1.12) to further close gaps in our genome obtained in step3.

NOTES

Hongling Zhou 06 Jun 2018

1-Note: using parameters'pregraph(-K 29 -p 20);contig(-M 2);map(-k 41);scaff(<default>)'

Hongling Zhou 06 Jun 2018

2-Note: using reads from all insert-size libraries

Repeat annotation_de novo

Step 4.

1. Run RepeatModeler(1.0.8) and LTR_FINDER(1.0.6), respectively, 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 (version 3.3.0) to find and then classify the repetitive sequences.

NOTES

Hongling Zhou 06 Jun 2018

1-Note: <default>

Hongling Zhou 06 Jun 2018

2-Note: using parameters '-nolow -no_is -norna -parallel 1'

Repeat annotation_homolog

Step 5.

Run RepeatMasker and ProteinMask (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

Hongling Zhou 06 Jun 2018

Using parameters 'RepeatMasker(-nolow -no_is -norna -parallel 1)ProteinMask(-noLowSimple -pvalue 0.0001)'

Gene prediction_de novo

Step 6.

Run Augustus (version 3.0.3) and GlimmerHMM (version 3.0.1) to de novo predict genes in the repeat-masked genome sequences.

NOTES

Hongling Zhou 06 Jun 2018

Using parameters 'Augustus(--species=zebrafish --uniqueGeneId=true --noInFrameStop=true --gff3=on --strand=both)GlimmerHMM(-d zebrafish -f -g)'

Using parameters '-d zebrafish -f -g'

Gene prediction_homolog

Step 7.

Download protein sequences of homlog species (danio rerio(release-64), gadus morhua(release-65), gasterosteus aculeatus(release-64), oryzias latipes(release-64), takifugu rubripes(release-64), and tetraodon nigroviridis(release-64)), 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

Hongling Zhou 06 Jun 2018

Using parameters '--min_align_coverage 0.3 --max divergence rate 0.3 --extend length for both sides of regions 2000'

Gene prediction_glean

Step 8.

Integrate genes predicted in step 8-9 to obtain the consensus gene set by using GLEAN.

NOTES

Hongling Zhou 06 Jun 2018

Filtering with criterion 'overlap cutoff 0.8 and at least one homolog support'

Gene prediction_adding RNA-seq

Step 9.

Perform TopHat (version 2.1.0) with default parameters to align filtered RNA-seq reads against gene set mentioned in Step10, 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

Hongling Zhou 06 Jun 2018

Filtering RNA sequencing data by SOAPnuke with parameters '-l 10 -q 0.5 -n 0.01 -Q 2”

Estimation of completeness

Step 10.

Run BUSCO(version 3.0.1) and map final gene set and genome to actinopterygii reference to assess the completeness.

NOTES

Hongling Zhou 06 Jun 2018

Using parameters '-e 0.001 -limit 3'