FEB 09, 2023

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.4r3l27ez4g1y/v1

Protocol Citation: Boyang Liu, Liangyu Cui, Zhangwen Deng, Yue Ma, Diancheng Yang, Yanan Gong, Yanchun Xu, Shuhui Yang, Song Huang 2023. The annotation pipeline for the genome of a snake. **protocols.io** https://dx.doi.org/10.17504/protocols.io.4r3l27ez4g1y/v1

Protocol status: Working

Created: Feb 06, 2023

Last Modified: Feb 09, 2023

PROTOCOL integer ID: 76443

The annotation pipeline for the genome of a snake

Forked from Fish genome assembly and annotation pipeline

In 1 collection

Boyang Liu[1], Liangyu Cui[1], Zhangwen Deng[2], Yue Ma[1], Diancheng Yang[3,4], Yanan Gong[3,4], Yanchun Xu[1], Shuhui Yang[1], Song Huang[3,4]

[1]College of Wildlife and Protected Area, Northeast Forestry University, Harbin 150040, China;

[2]Guangxi Forest Inventory and Planning Institute, Nanning 530011, China;

[3]Anhui Province Key Laboratory of the Conservation and Exploitation of Biological Resource, College of Life Sciences, Anhui Normal University, Wuhu 241000, China;

[4]Huangshan Noah Biodiversity Institute, Huangshan 245000, China

GigaScience Press BGI

博洋 Boyang Liu

ABSTRACT

Here are detailed methods use for the annotation of various snake genomes.

Repeat annotation_de novo

1 1) Run RepeatModeler to build a *de novo* library based on the input assembled genome sequence.
2) Using the library constructed in step 5 as the 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

2 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 the Repbase library (v. 17.01).

Note

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

Gene prediction_preparation

3 Mask these repetitive regions obtained above (step 4-6) with 'N's.

Note

Before gene prediction, mask the TE's (transposable elements) in the genome.

Gene prediction_de novo

4 Run Augustus (v3.0.3) to *de novo* predict genes in the repeat-masked genome sequences.

Note

using parameters "--species=Ophiophagus_hannah --uniqueGeneId=true --noInFrameStop=true --gff3=on --strand=both" when running Augustus.

Gene prediction_homolog

5 Download the publicly available protein sequences of representative homologous snake species, align these against our masked genome sequences with BLAT, and then based on the BLAT mapping results, GeneWise (v2.4.1) is then run to predict the genes.

Gene prediction_transcriptome

6 Then filter RNA-seq data using Trimmomatic(v0.30). The resulting data is then assembled by Trinity (v2.13.2). PASA(v2.0.2) was finally used to align transcript against the snake genome of interest to obtain gene structures.

Note

default parameters

Final gene set_MAKER

7 Integrate the genes predicted in step 4-6 to obtain the consensus gene set using the MAKER pipeline (v3.01.03).

Functional annotation

8 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 (v5.52-86.0) with seven different models (Profilescan, blastprodom, HmmSmart, HmmPanther, HmmPfam, FPrintScan and Pattern-Scan)