Gene modeling and prediction

Hansheng Zhao

Abstract

We performed a considerate prediction of intact protein-coding gene models using three independent approaches, i.e.,de novoprediction, homology-based method, and RNA-Seq approach.

Citation: Hansheng Zhao Gene modeling and prediction. **protocols.io** dx.doi.org/10.17504/protocols.io.rf7d3rn

Published: 03 Jul 2018

Protocol

De novo prediction

Step 1.

The repeat masked assembly was firstly annotated by AUGUSTUS (version 3.3) with default parameters, which was a de novopredictor based on the self-trained model.

SOFTWARE PACKAGE (LINUX -)

AUGUSTUS, 3.3

Homology-based prediction

Step 2.

In the homology-based prediction, we used seven species as reference datasets,i.e., Elaeis guineensis, *Phoenix dactylifera*, Brachypodium distachyon, Oryza sativa, Setaria italic, Sorghum bicolor, and Zea mays (see Availability of supporting data for individual genome version). Their protein sequences were downloaded for ENSEMBL database [30]and were aligned to the *C. simplicifolius*and *D. jenkinsiana*assembly using TBLASTN (version 2.2.26) with an E-value cutoff of 1e-5, respectively. Then, the splicing patterns were generated by GeneWise (version 2.0).

SOFTWARE PACKAGE (LINUX -)

GeneWise, 2.0

RNA-Seq analysis

Step 3.

In the RNA-Seq analysis, HISAT2 (version 2.0.2) was used to identify exon-intron splicing junctions and refine the alignment of the RNA-Seq reads to the genome. Then, we used Cufflinks (version 2.2.1) to define some protein-coding gene models in C. simplicifoliusand D. jenkinsiana, respectively

SOFTWARE PACKAGE (LINUX -)

Cufflinks, 2.2.1

Intergration

Step 4.

We integrated the evidences from the three above independent predictions using MARKER (version 2).

SOFTWARE PACKAGE (LINUX -)

MAKER, 2