Developing A Computational Pipeline For High-throughput Sequencing Data In Plant Genomes

Yiran Zhang, Swaminathan Padmapiya, A. Fennell, Qin Ma
Plant Science Department and Computer Science, South Dakota State University, Brookings, SD, 57007; 1Plant Science Department, South Dakota State University, Brookings, SD, 57007; 2Computational Systems Biology Lab, South Dakota State University, Brookings, SD, 57007; 3College of Computer Science and Technology, and School of Public Health, Jilin University, Changchun, China; 4BioEnergy Science Center (BESC), US Department of Energy, Oak Ridge, Tennessee 37831; 5College of Computer Science and Technology, and School of Public Health, Jilin University, Changchun, China.

Introduction

- Developments of high-throughput sequencing biotechnologies enable monitoring thousands of molecules (RNA, protein or metabolites) simultaneously.
- These technologies have the potential to revolutionize how we study plants and other organisms.
- RNA-sequencing (RNA-seq) is a common and important high-throughput sequencing technology. However, mapping uncertainty exists and becomes the computational bottleneck in RNA-seq analysis.

Objective

- Design an RNA-seq data analysis pipeline and corresponding computational tools/programs in a plant genome.
- Break the computational bottleneck in RNA-seq analysis.

Methods

Predict relative weights among a gene and its homologs under a certain condition, based on large-scale co-expression modules in a target plant genome. Then the predicted weights can help assign those multiple mapped RNA reads.

Key Work

We are using co-expression information from an independent microarray study as a training set to predict relative weights among a gene and its homologs under a certain condition for use in RNA-seq analysis studies. Co-expressed genes with unique mapped reads will be set as a reference to guide assignment of the multiple mapped reads base on a well defined mathematical model.

Technology and tools

Check the per base sequence quality and produce a report including k-mer content, GC content, adapter content and so on.

Discard the bad-quality reads before going to the next step.

"Unique mapping" means a read will be directly discarded if it can be mapped to different genomic locations with the same matching scores.

"Even distributing multiple mapped read". This strategy allows these ambiguous reads to be evenly distributed among multiple mapped genomic locations having the same matching scores.

The overall expression correlation among different replicate samples will be conduct using the read counts for all the genes in a target genome.

Overall correlations among different replicates are expected to be highly similar, otherwise additional computational efforts are required to reduce the batch effect.

Trinity is the choice for de-novo transcript assembly. It is noteworthy that these assembly tools are designed and evaluated based on human/mouse genome.

Relevant parameters will be optimized in our pipeline for plant genomes.

For all the assembled transcripts, the biological pathway enrichment analysis will be conduct using the web server, DAVID, in support of further biological interpretation or hypothesis test.

RNA-sequencing reads

Read quality check

Qualified reads mapping

Replicate sample quality check

Gene assembly

Clis-regulatory motif identification

Differential expression analysis

Pathway enrichment analysis

Database of the temperature response genes

GSEA, DAVID

See the enrichment analysis the raw variant associations to combining all raw variants

References

2. An RNA-seq analysis pipeline for switchgrass integrating co-expression network analysis, Qin Ma1,2,4, Xin Chan2,4,5, Xin Shun Ding3,4, Youhong Tang1,4, Rick Nelson3,4 and Ying Xu2,4,5
1Plant Science Department, South Dakota State University, Brookings, SD, 57007; 2Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, GA, USA; 3Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma 73401, USA; 4BioEnergy Science Center (BESC), US Department of Energy, Oak Ridge, Tennessee 37831; 5College of Computer Science and Technology, and School of Public Health, Jilin University, Changchun, China;