





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

Yiran Zhang, Swaminathan Padmapriya, 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; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State University, Brookings, SD, 57007; Plant Science Department, South Dakota State Un

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 largescale 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 Main Pipeline Results Summary **RNA-sequencing reads** Check the per base sequence quality and . 📀 Basic Statistics FastQC, is used produce a report including k-mer content, GC . O<u>Per base sequence quality</u> Basic Statistics to do quality content, adapter content and so on • OPer tile sequence quality control checks for . OPer sequence quality scores Value Measure the large-scale 1709_6079_5797_N_FSH_2_CGATGT_R1.fastq Filename er base sequence content File type Conventional base calls raw data, i.e., all • O<u>Per sequence GC content</u> **Read quality check** Sanger / Illumina 1.9 Encoding FastX-FastQ FastQC Total Sequences 15181252 . Ø<u>Per base N content</u> the sequenced Discard the bad-quality reads before going to Sequences flagged as poor qualit • OSequence Length Distribution RNA reads. Sequence length the next step Sequence Duplication Levels • Overrepresented sequences "Unique mapping" means a read will be directly . 📀 <u>Adapter Content</u> **Qualified reads mapping** HISAT2 discarded if it can be mapped to different . 🐼 <u>Kmer Conten</u>t Two different mapping genomic locations with the same matching strategies will be considered scores. for the further analyses regarding how to deal with Quality scores across all bases (Sanger / Illumina 1.9 encoding "Even distributing multiple mapped read". This the multiple mapped reads ╺┥┥┥┥┙┥┥┥┥┥┥┥┥



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;

