

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;

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

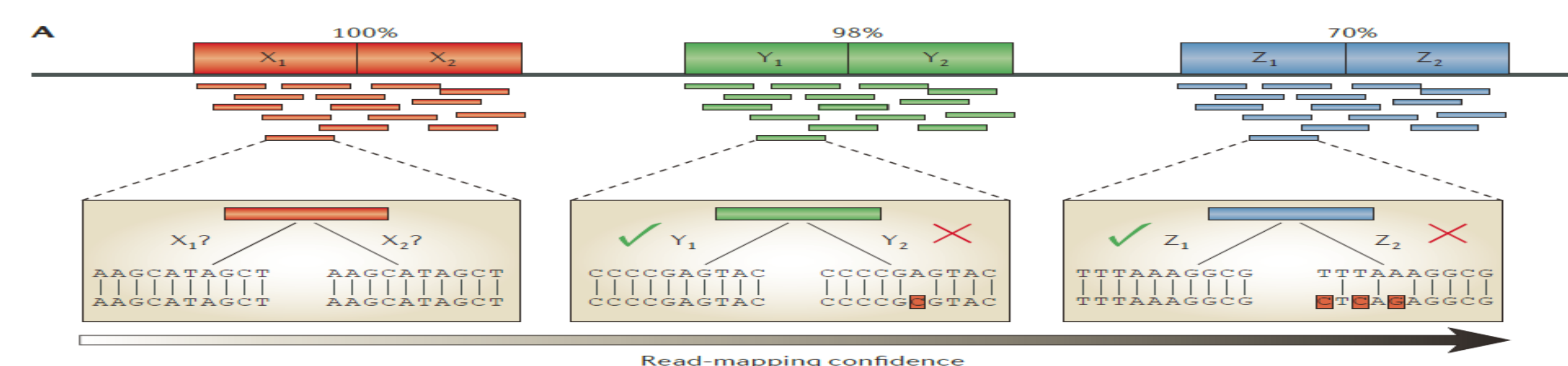
- Design an RNA-seq data analysis pipeline and corresponding computational tools/programs in a plant genome
- Break the computational bottleneck in RNA-seq analysis

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.

Methods

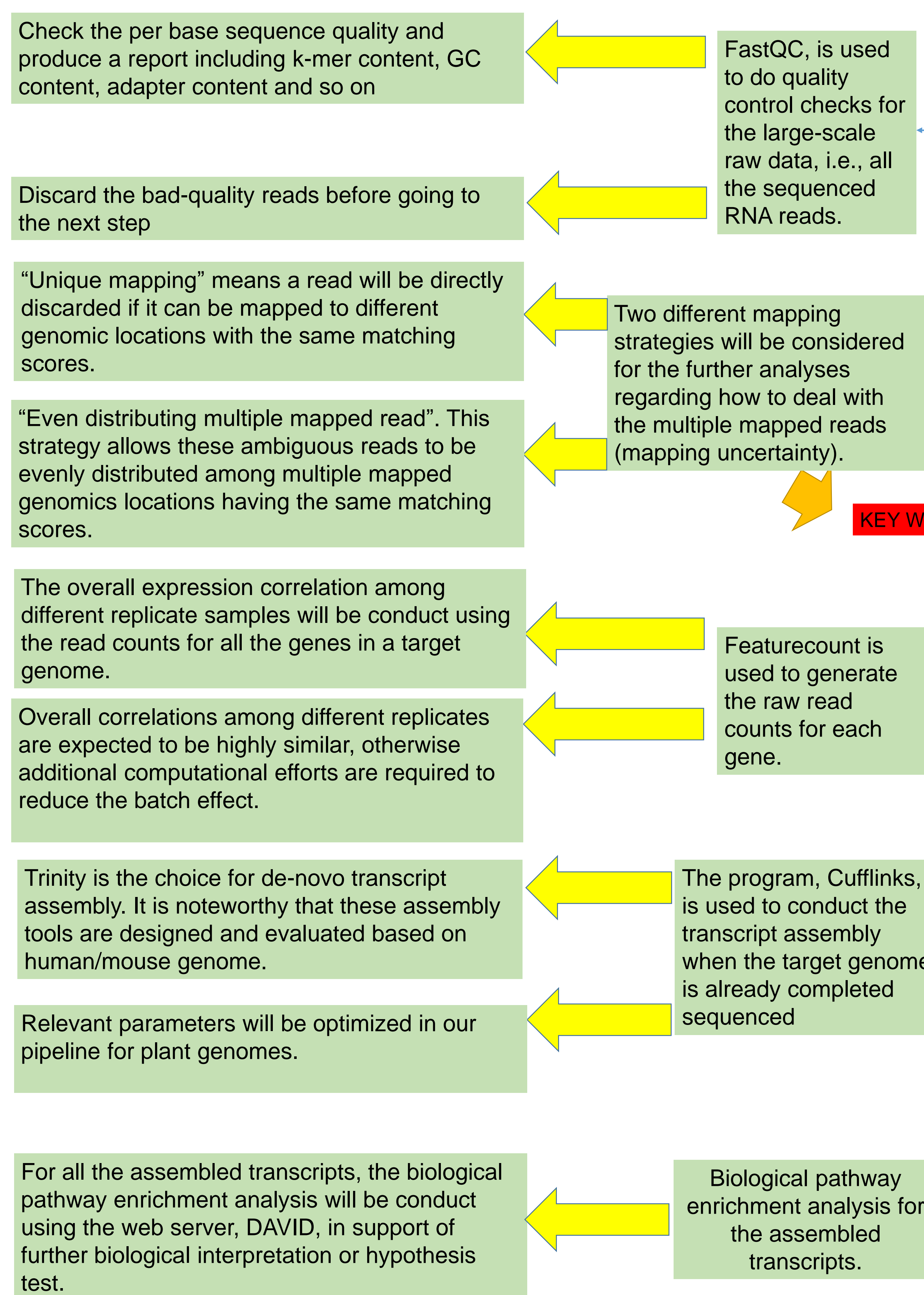
Predict relative weights among a gene and its homologs under a certain condition, based on large-scale co-expression modules in a target plant genome. Then the predicted weights can help assign those multiple mapped RNA reads



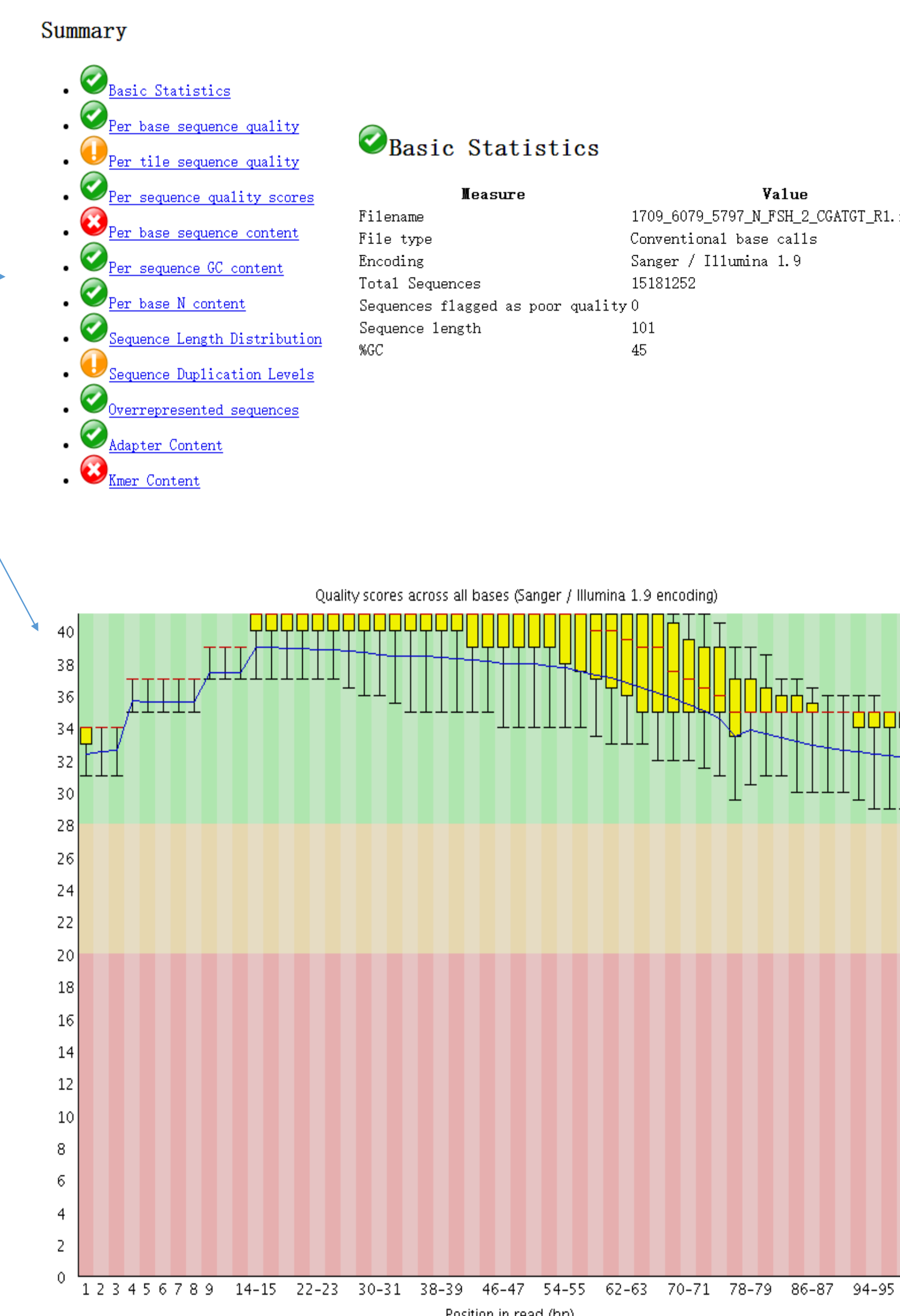
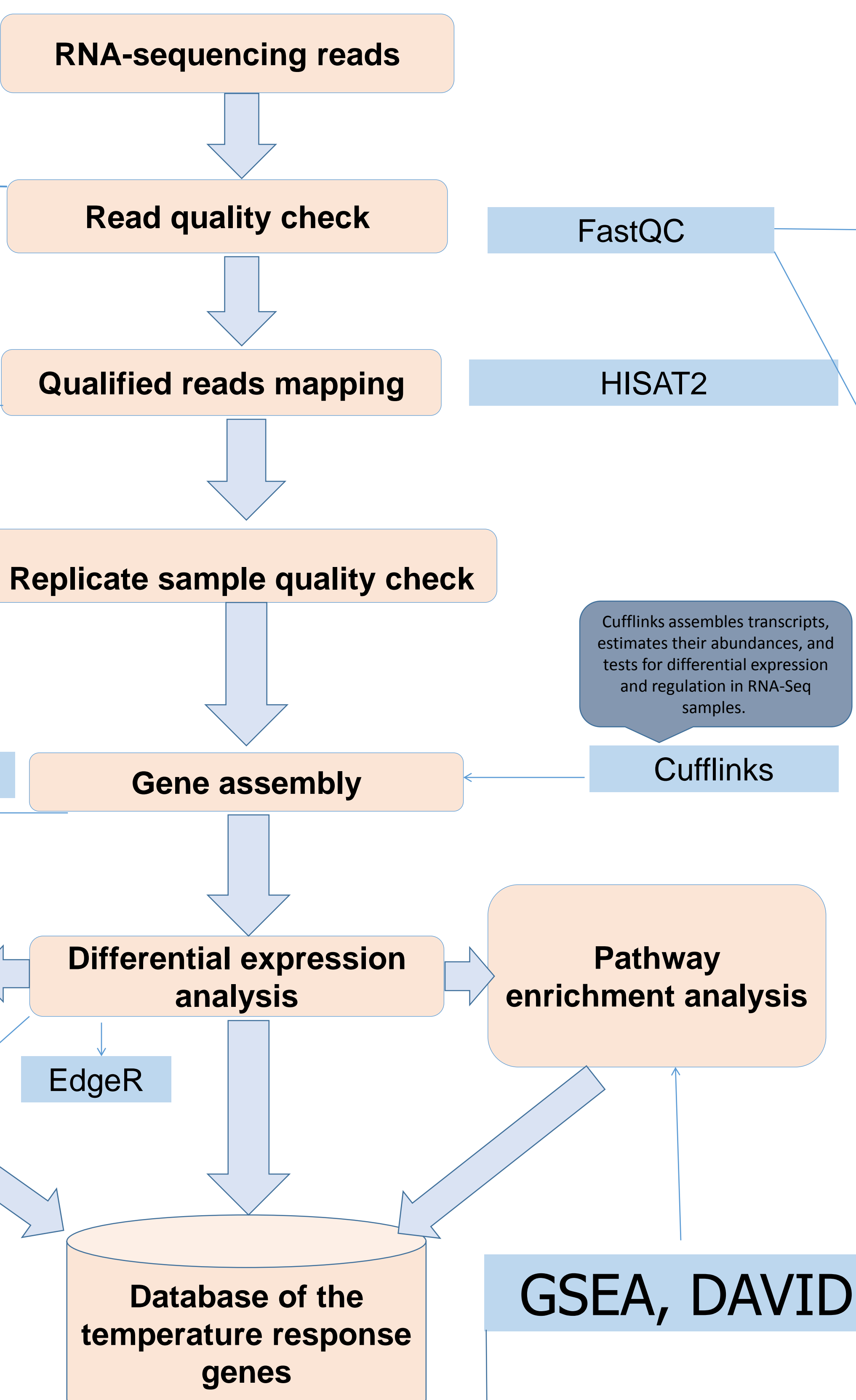
Technology and tools

Main Pipeline

Results



KEY WORK



References

1. "South Dakota State University." *South Dakota State University*. N.p., n.d. Web. <<https://www.sdstate.edu/ps/research/bioinformatics/index.cfm>>.
2. An RNA-seq analysis pipeline for switchgrass integrating co-expression network analysis, Qin Ma^{1,2,4}, Xin Chen^{2,4,5}, Xin Shun Ding^{3,4}, Yuhong Tang^{3,4}, Rick Nelson^{3,4} and Ying Xu^{2,4,5}
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;

gene set enrichment analysis: Do rare variant association test statistics by combining all rare variants