NEBRASKA CENTER FOR ENERGY SCIENCES RESEARCH

Developing a software tool to identify genotype-specific RNA splicing events in rice, maize, and sorghum populations to link them to the phenotype of oil accumulation and biomass

Yashraj Purbey[1], Bin Yu[2], Chi Zhang[2]

[1]School of Computer Science and Engineering, College of Arts and Sciences, UNL, 68588
[2]School of Biological Sciences, Department of Biochemistry, Center of Plant Science Innovation, UNL, 68588

BACKGROUND

Pre-mRNA splicing is an essential step for the regulation of gene expression. The removal of introns and joining exons during the mRNA maturation process is an essential step in gene regulation for biological processes in most eukaryotes. Alternative splicing (AS) is an important post-transcriptional regulatory mechanism in plant response to abiotic stresses (Laloum et al., 2018). There are five basic types of AS: alternative donor site (AltD), alternative acceptor site (AltA), alternative position (AltP), exon skipping (ExonS), and intron retention (IntronR). Previously, we developed a software tool to quantify and visualize Variations of Splicing in Population (VaSP). VaSP can quantify splicing variants from short-read RNA-seq datasets and discover genotype-specific splicing (GSS) events, which can be used to prioritize causal pre-mRNA splicing event in GWAS. During this summer, we upgraded the software package by developing new scoring function, and new function to identify different types of alternative splicing in plant populations, including AltD and AltA, which were missing in the first version.

We developed several new functions and upgraded the software package to VaSP2. We applied VaSP2 on a rice, maize, and sorghum populations, using a large set of short read RNA-seq samples. We used VaSP2 to discover new types of GSS events, which can prioritize candidate causal pre-mRNA splicing events for the associate study. For these GSS events, we conducted association studies to link them to the phenotype, such as oil accumulation and biomass. The preliminary study showed promising result. We plan to publish the new version of the software package, VaSP2, in this Fall Semester.

METHODS

- We collected three datasets for rice, maize, and sorghum populations. Each dataset has more than 300 RNA-seq data for different genotypes and conditions, including various stresses and planning conditions. The numbers of samples for every species are listed in the following table.

#	Species	# of samples in the population
1	Rice	289
2	Maize	365
3	Sorghum	337

- VaSP could be download from https://bioconductor.org/packages/devel/bioc/html/VaSP.html
- The new version of scoring function for quantifying splicing strength is called 3Sv2, Single Splicing Strength (3S) score: $3S = \frac{R}{\sum C_i}$, where R is the count of junction-reads for a single intron, and Ci is the average coverage of the i-th isoform for the parent gene. This new scoring function was implemented as an R function S3v2().
- This function needs to read RNA-seq bam files and collect read alignment information for all 300 samples in a population. We carefully designed the function to make it efficient and fast.
- Our methodology for analyzing alternative splicing events involved three main stages: identification of comparable splicing events, quantification of splicing score, and clustering of samples.

METHODS (CONT.)

1. Identification of Alternative Splicing Events

To analyze differential splicing, we first identified introns that could be directly compared. We processed the RNA-seq junction reads to find intron clusters. Following the principles of the Leafcutter method, introns that share a common splice start or end site were grouped into intron clusters. An intron cluster is defined as a group of two or more introns that share a common splice site - either (the "start") or (the "end"). This approach allows for the identification and quantification of splicing events without relying on pre-existing gene annotations and allows us to directly compare the splicing events originating from the same genomic location.

2. Quantification of Splicing Ratios

For each intron cluster, we quantified the usage of the alternative splice junctions across all samples. This was done by calculating a splicing score for each of the two introns within a cluster. The score represents the proportion of reads supporting one splicing event relative to the total reads covering the gene. For a given intron cluster with two junctions, this process yields two scores per sample. These two scores were then used as the x and y coordinates for the subsequent clustering analysis.

3. K-Means Clustering of Splicing Patterns

To identify distinct splicing patterns across the samples, we used K-means clustering. The two splicing scores calculated for each sample were plotted in a 2D space. This algorithm looks at the scores for every sample and automatically sorts them into a pre-defined number of groups (in this case, two). The goal is to put samples with the most similar splicing scores into the same group.

- To identify GSS events associated with phenotype, a linear model was used to fit the phenotype with the genotype as follows: y ~ x + PC1 + PC2 + PC3 + PC4, where y is the phenotype, x is the clusters of each GSS event (genotype), PC1-PC4 are the first four principle components from clusters of all GSS events under salt stress condition (control population structure).

RESULTS

The splicing model of AltD and AltA is shown in the following figure.

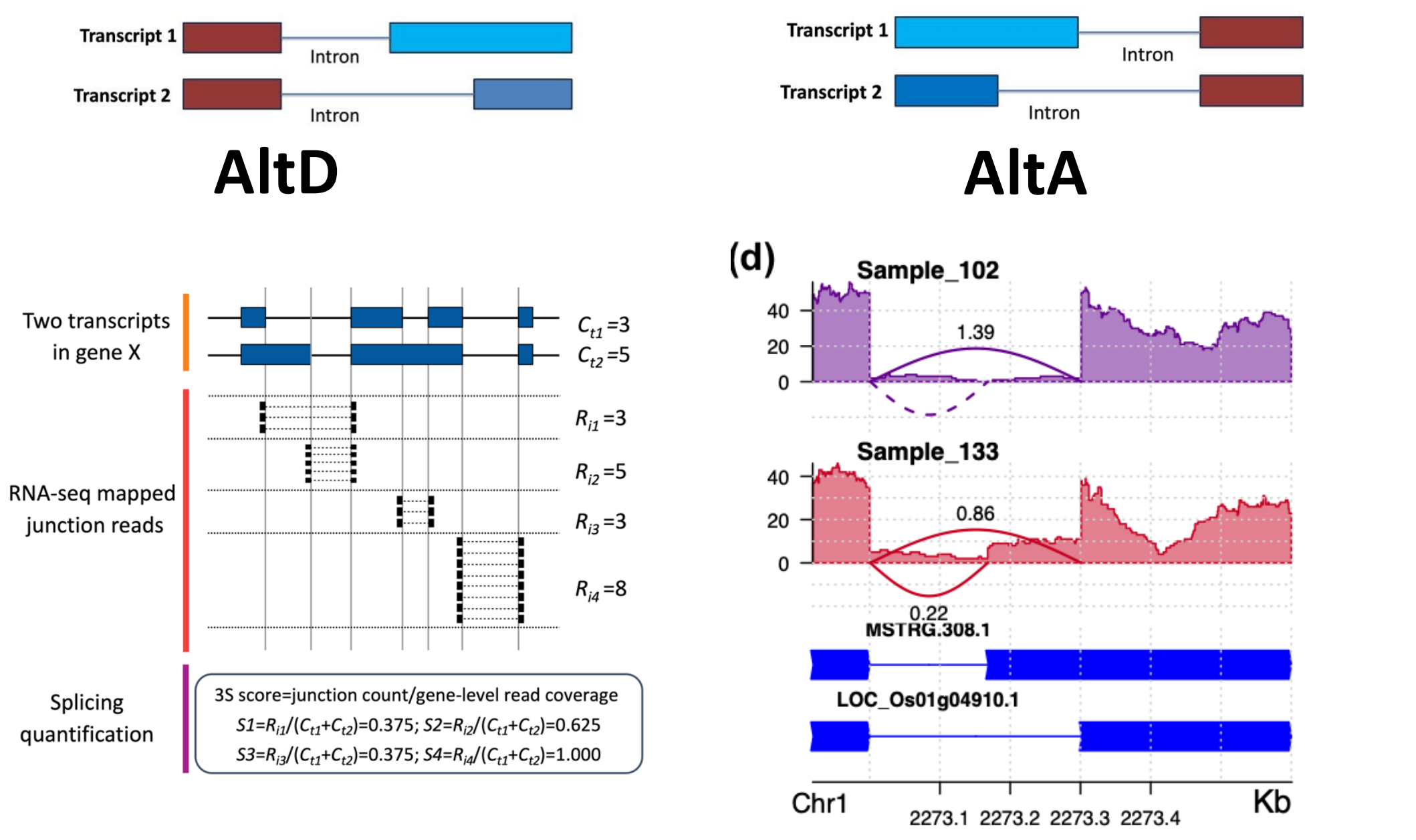

Figure 1. Left figure shows the way to estimates 3S scores based on junction read counts normalized by gene-level read coverage. Right Figure is the visualization of differential splicing region of the gene MSTRG.183 with splicing scores displaying.

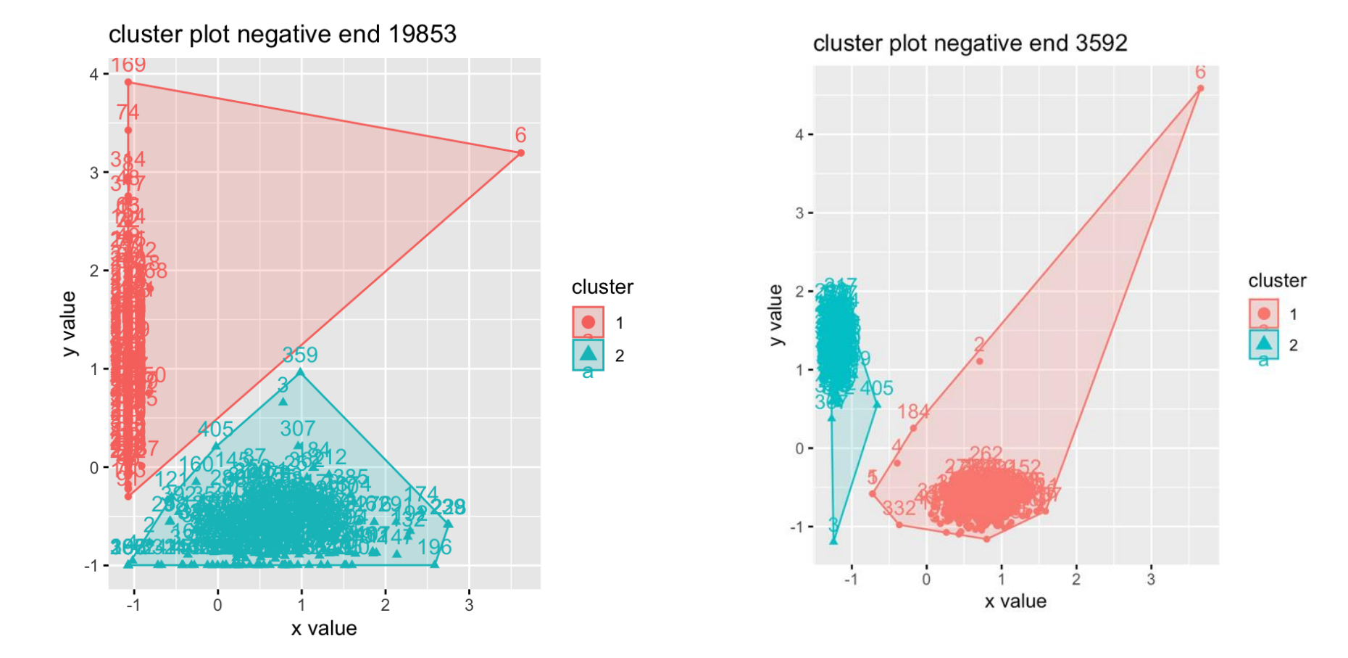

Figure 2. Two example of clusteres that are a GSS AltT or AltD events from rice population;

RESULTS

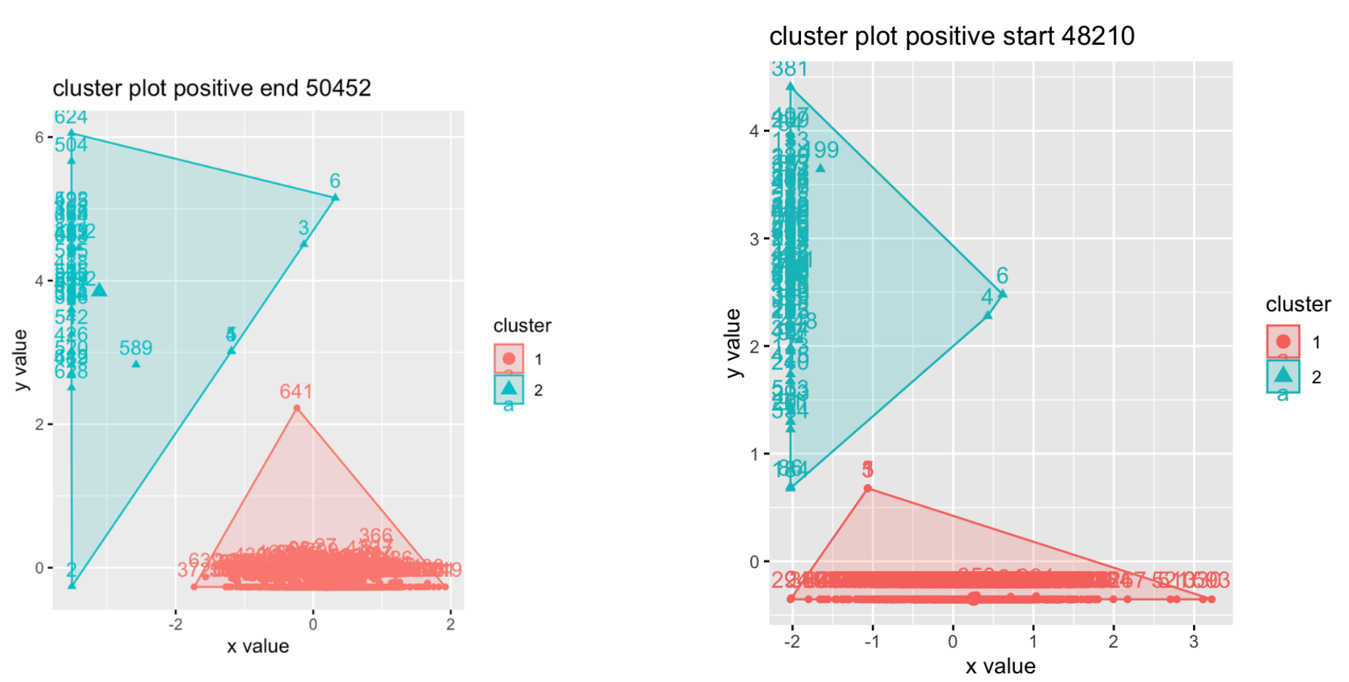

Figure 3. Two examples of clusters from maize population. One figure indicates to a GSS AltT or AltD event.

We tested the new function to identify GSS AltA or AltD events in rice, maize, and sorghum populations. In each population, there are more than 300 samples. Usually, it took several hours for our software to calculate the 3S scores and identify GSS AltA or AltD events. The following table shows the total numbers of GSS AltA or AltD events that we discovered in rice, maize, and sorghum. Since we are still working on sorghum data, the numbers showed here are not the final result.

	AltD	AltA
rice	96	128
maize	118	234
sorghum	77 (E)	103(E)

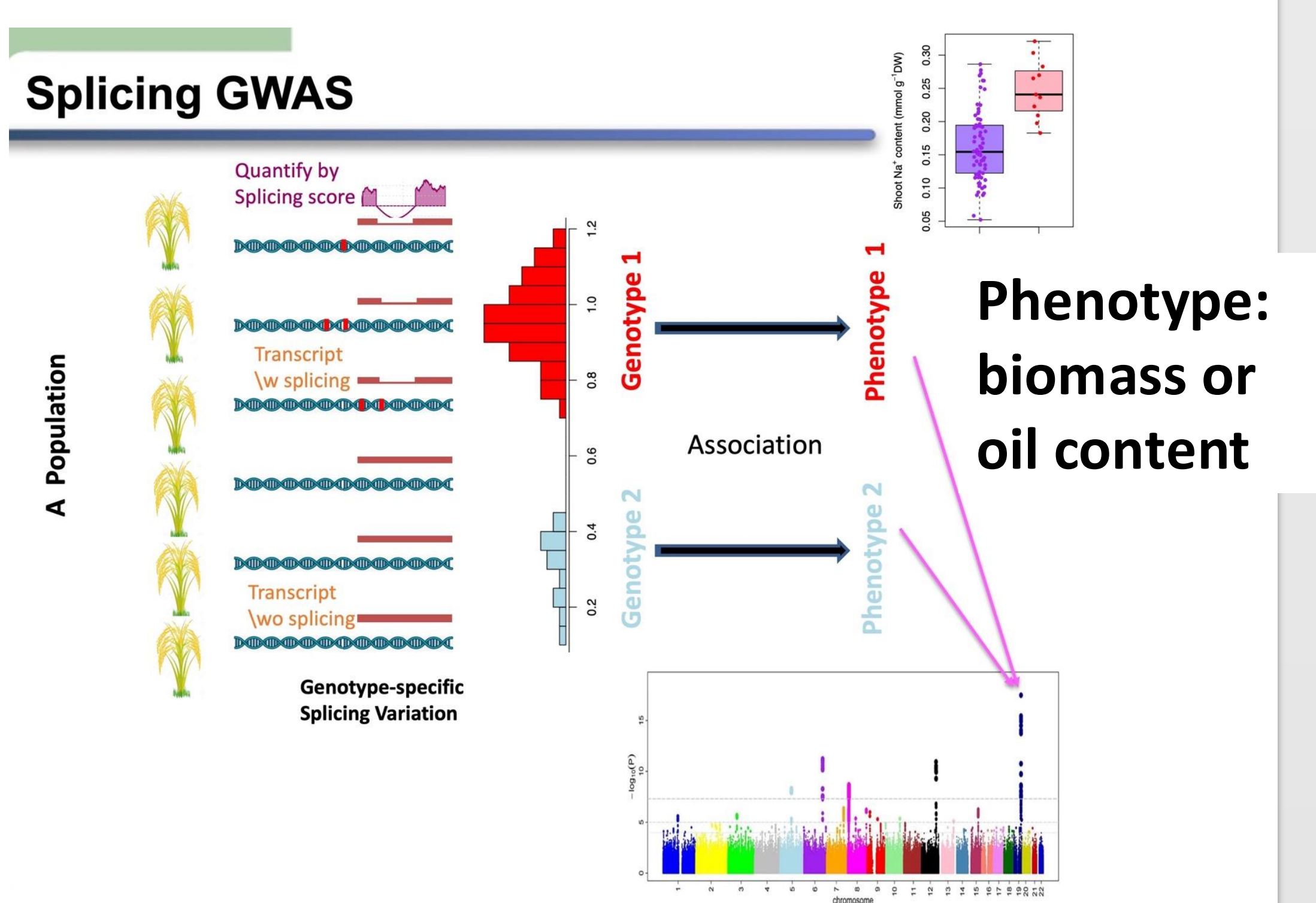

Figure 4. The scheme for us is to conduct association study to link GSS events to phenotype.

CONCLUSION

We developed a software package to study RNA alternative splicing in plant populations. We applied this tool to rice, maize, and sorghum populations. We explored how to link genes change & adapt over time. These analysis can help us understand the intricate relations that each splicing event has on the overall plants and corresponding phenotype. This can help us construct better plant lines that have an increased oil yield or biomass.

REF/ACKNOWLEDGEMENTS

1. This work was directly supported by the Nebraska Public Power District through the Nebraska Center for Energy Sciences Research (NCESR) at the University of Nebraska-Lincoln.
2. This work also partially supported by DOE.

3. H. Yu, Q. Du, M. Compbell, B. Yu, H. Walia, Chi Zhang*. Genome-Wide Discovery of Natural Variation in Pre-mRNA Splicing and Prioritizing Causal Alternative Splicing to Salt Stress Response in Rice. New Phytologist (2021); 230: 1273-1287
4. https://bioconductor.org/packages/devel/bioc/html/VaSP.html