DEVELOPMENT OF A MULTI-OMICS APPROACH TO IDENTIFY HIGHLY CORRELATED TRANSCRIPTOMIC, PROTEOMIC AND METABOLOMIC SIGNATURES IN MAIZE B73 AND FR697 DROUGHT STRESSED NODAL ROOTS

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By SIDHARTH SEN

Dr. Trupti Joshi, Dissertation Supervisor

MAY 2021

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

DEVELOPMENT OF A MULTI-OMICS APPROACH TO IDENTIFY HIGHLY CORRELATED TRANSCRIPTOMIC, PROTEOMIC AND METABOLOMIC SIGNATURES IN MAIZE B73 AND FR697 DROUGHT STRESSED NODAL ROOTS

presented by Sidharth Sen, a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Trupti Joshi

Dr. Chi-Ren Shyu

Dr. Dong Xu

Dr. Melvin J. Oliver

Dedication

To my parents, Kiriti and Devika Sen, for their unconditional love, support, and encouragement.

ACKNOWLEDGEMENTS

First, I would like to thank my adviser, Dr. Trupti Joshi. Her guidance, encouragement, patience, and continuous support over the duration of my research have been invaluable, without which, the completion of this dissertation would not have been possible.

I would like to thank my committee members – Dr. Chi-Ren Shyu for not only his advice which helped me maintain focus during my Ph.D. training, but also the inclusion in various social activities which enriched my time here. Dr. Dong Xu for his advice and exposing me to various research areas through the DigBio lab. Dr. Melvin J. Oliver for aiding in critical thinking and fun discussions related to maize genetics.

I additionally thank my colleagues and informal advisors from the roots in drought team - Dr. Robert E. Sharp, Dr. David M. Braun, Dr. Felix B. Fristchi, Dr. Scott C Peck, Professor Jonathan T. Stemmle, Shannon K. King, Tyler McCubbin, Dr. Laura A Greeley, Cheyenne Becker, Dr. Rachel Mertz. They not only collected the data used for my research, but also introduced me to the various methods and techniques used in an agricultural laboratory, giving me the appreciation for the effort required to generate high quality samples. My research and the entire group are supported by the NSF grant IOS- 1444448

I am also thankful to the current and past members of the Trupti Joshi lab for their friendship and collaborations – Shuai Zeng, Dr. Sadia Akhtar, Dr. Juexin Wang and Zhen Lyu.

I am also thankful to the DGS - Dr. Timothy Matisziw for keeping me on track for my academic progress. Tracy Pickens with her on point guidance and the epic Robert Sanders

for his stress-free help with administrative matters were which were instrumental to my progress.

Last, but not least, I would like to thank my parents Kiriti and Devika Sen, for their undying support during not only my Ph.D. but also during all my previous academic and professional endeavors and in instilling in me the ethics and values which has helped me overcome many hurdles and succeed in life.

Table of Contents

List of Tables

List of Figures

LIST OF EQUATIONS

List of Abbreviations used:

PCR – Polymerase Chain Reaction

sGCCA - sparse Generalized Canonical Correlation Analysis

PLS - Partial Least Square

PCA - Principal component analysis

WGCNA - weighted gene co-expression network analysis

DE – Differential Expression

KEGG - Kyoto Encyclopedia of Genes and Genomes

DNA - Deoxyribonucleic acid

cDNA – complementary DNA

RNA-Seq - ribonucleic acid - sequencing

miRNA – microRNA

piRNA – piwi-interacting RNA

STRINGdb - Search Tool for the Retrieval of Interacting proteins database

STITCHdb - Search Tool for Interactions of Chemicals Database

GDD - Growing Degree Days

MPa – Mega Pascal

RP/UPLC - reversed-phase ultra-high-performance chromatography

HILIC/UPLC - Hydrophilic interaction ultra-performance liquid chromatography

MS/MS - Tandem mass spectrometry

PFPA - Perfluoropentanoic Acid

FA - Formic Acid

LIMS – Laboratory Information Management System

DAP – Degree after Pollination

CCS – Circular Consensus Sequence

SAGE - Serial analysis of gene expression

GO – Gene Ontology

Dissertation Structure

This dissertation is organized into 7 chapter. Chapter 1 introduces overarching goal and background of this project, along with a brief overview of the informatics methods used to explore the objective. Chapter 2 presents the data used in this dissertation, including how they were collected and processed. Chapter 3 presents the transcriptomics analysis workflow along with an exploration of the results. Chapter 4 introduces the multiomics strategy, using 3 omics datasets from the FR697 datasets and exploration of the results. Chapter 5 details the generation of the FR697 specific transcriptome. Chapter 6 introduces the development of interactive visualization methods for multiomics datasets. Chapter 7 concludes this dissertation and presents the way forward and future directions.

Abstract

Maize is one of the most important crops grown in the continental US and worldwide, and as such, major interest is directed towards understanding the impact of drought conditions on maize growth and development. Nodal roots, which develop from the base of the stem and produce the framework of the mature root system, can continue to grow under water stress conditions that inhibit the growth of the leaves and stem. To better understand the molecular mechanisms that led to this remarkable ability, we analyzed multiomics (transcriptome, proteome, metabolome) datasets generated from the growth zone of nodal roots collected from the reference inbred line B73 and from inbred line FR697, which exhibits a relatively greater ability to maintain root elongation under water-stressed conditions. We developed an informatics analytics pipeline consisting of a discriminatory multiomics data integration approach combining sparse Generalized Canonical Correlation Analysis (sGCCA) and generalized Partial Least Square analysis (PLS) to incorporate all datasets into one holistic global network and form clusters spanning all omics levels. Significant elements from these clusters were connected to various observations associated with water stress in the root tip samples and reinforced by their roles in biological pathways. We also generated an annotated "SuperTranscriptome" assembly from Pacbio Iso-Seq and RNA-Seq datasets to serve as a representative assembly for the FR697 genotype. The results were incorporated into the KBCommons maize database for storage and analysis from various viewpoints. To visualize interactions between the many elements, we are also developing a suite of 3D visualization, collectively called the "KBCommons Omics Studio", integrated with the KBCommons framework. Using these methods, we showcase possible biomarkers related to drought stress and allied observations. Supported by NSF Plant Genome Program IOS #1444448.

Chapter 1: Introduction & Motivation

1.1 Effects of drought on the maize plant

Maize (*Zea mays* L.) is considered one of the most important food crops in the world, utilized not only as a food crop but also as feed for livestock, to produce various chemicals and to generate biofuels. Maize is grown in various countries, with up to 90 million acres of arable land utilized for maize cultivation in the United States alone[1]. In 2012, almost 78% of maize growing areas in the US experienced drought conditions and in subsequent years, large regions have continued to face substantial drought events, resulting in lower-than-expected harvest yields. The socio-economic costs of drought are well recognized, and various organizations closely monitor its effects on food shortages both country- and world-wide[2].

As such, there is a major interest in understanding the effects of drought on the maize plant, in particular the growth and functioning of the root system. The maize plant has a distinct root system, comprising of seminal roots which form during the germination phase; and nodal roots, which develop from the base of the stem and produce the framework of the mature root system, thus providing the bulk of water and nutrient uptake needed for the mature plants. Early studies reported that the nodal roots[3], can continue to grow under water stress conditions that inhibit the growth of the leaves and stem[4], [5]. This phenotype indicates an innate survival mechanism present in the maize plant. However very little is known about the underlying biological mechanisms and genetics associated with ability.

In plants, organ growth encompasses two types of cellular activity – cell division and cell expansion[6]. In primary and nodal roots of maize, cell division occurs in the apical 3 mm, whereas cell elongation occurs throughout the apical 10 mm[7]. The dynamics of both of these activities are altered in roots exposed to water stress[8]. Thus, to gain insights into the genetic mechanisms behind drought stress adaption or acclimation in maize nodal roots, apical 1-cm samples of the root tips were collected from B73 genotype plants grown under well-watered and water-stressed conditions in the field by the drought root team lead by the Sharp lab. Similar samples were also collected from maize inbred line FR697 plants grown alongside the B73 plants. Originally developed by Illinois Foundation Seeds Inc., the FR697 genotype exhibits a greater ability for root growth maintenance when compared to the reference inbred line B73 under similar water stress conditions[9]–[11].

One issue while collecting samples from field grown plants is the effect of various environmental factors upon them. Major variation and batch effects are caused by factors as simple as amount of sunlight during sample collection, soil hardness, previous day's or previous week's weather[12]–[16]. With so many variables in play, a lot of unnecessary noise is added to expression and/or quantitative data, resulting in various issues during data harmonization and integration. To alleviate some of these issues, FR697 maize nodal root tip samples were also collected from plants grown under drought stress in the Sharp lab using their recently developed method called the "split root growth chamber system" [17]. This system grows maize nodal root samples in a controlled environment and at precise water deficit levels, greatly reducing environmental effects. These root tip samples from both field and lab plants were used to generate multiple replicates of RNA-Seq datasets. The lab samples were also pooled to produce proteomics and metabolite datasets as well.

These datasets form the core of several data mining strategies to detect key elements related to drought response signatures presented in this dissertation.

1.2 Data mining strategies using omics datasets

The overarching goal of this project is to understand the biochemical mechanisms related to root growth maintenance and survival adaptation of the maize plant in drought stress conditions. One way to gain insights into the mechanisms of an organism is by calculating and comparing the abundance of transcripts per gene present in the genome of a species. This gives a snapshot of gene transcription activity of an organism by quantifying the expression levels for the samples under study. These datasets can also be compared with each for differential gene expression patterns, detecting genes which are significantly expressed in one group or another.

Significantly expressed genes can be organized into highly correlated clusters of interest, using unsupervised statistical tools like WGCNA[18] suggesting some sort of interaction or influence on each other. Genes coding for transcription factors within these correlated clusters can be highlighted and based upon protein interaction information taken from public databases such as STRINGdb[19], [20], be used to predict if any of the genes contribute to various observed phenotypes. Such results when overlayed over pathway information taken from public databases such as KEGG[21] or Reactome[22], has become a robust way to find unique interactions.

Another strategy is to look at omics datasets from the same bio samples. Traditionally, for such an “multiomics” approach, each omics dataset is analyzed on its own. Studies start with taking the first omics dataset in the series, generating a subset of significant elements,

and moving onto the next omics level, where a smaller set of connected significant elements are selected[23]–[25]. This results in a filter-funnel down approach with each layer constrained to elements which have the support of known connections. Thus, many important and related elements are usually left out, especially elements directly correlated to a diseased state or mutation state of individuals[26]. These left out elements could be from alternative pathways[27] or connected to control mechanisms which are either activated or de-activated due to the changes in the state of the individual; and thus, of great interest for further exploration.

Thus, the different omics layers need to be integrated and treated as larger global dataset, rather than split up into groups of local interactions, if we want to find these elements which act as nodes in the network of components. To integrate and analyze transcriptome, proteome and metabolite data, multiple strategies have been developed and used in the research community. Most integration methods focus on pairs of omics datasets[28], [29] such as transcriptomics and metabolomics, etc. Multivariate methods are usually the common feature detection method in such constrained cases, however where larger more complex datasets are to be analyzed; feature selection methods[30], [31] are employed to filter datasets down to key components.

1.3 De-novo Transcriptome assembly

While a relatively high-quality reference genome is available for the maize plant, it is based upon the B73 genotype. This leaves the door open for the presence of unique genes or formation of isoforms unique to the FR697 genotype. A de-novo transcriptome assembly using RNA-Seq short reads produces contigs which represent the transcripts formed during gene expression and can be used an effective substitute for a genome assembly. Tools like the Trinity assembler[32] use a large number high-quality RNA-Seq datasets such as those from the FR697 field and lab samples to generate a collection of contigs. Many of these contigs might be repeats of the same expressed transcripts and are usually reduced by using dedicated clustering algorithms such as CD-HIT[33]. To confirm that these transcripts represent all or most of the expressed genes from the samples, they are annotated by predicting their translated proteins and matching them with a database of known proteins. Using an expanded proteome such as combination of multiple plant species or even an entire kingdom is a viable strategy to annotate unrecognized transcripts.

1.4 Data Visualization

One key aspect of such informatics studies is the reporting of results in an abstract but informative manner. Most results are reported in the form of 2 dimensional graphs[34]–[37] such as PCA plots, Venn diagrams, heatmaps, etc. Since most of the observations in omics studies are expression based, PCA or correlation plots can visualize most of the observations as logical categories over multiple components. This however is severely limited in the context of systems biology and multiomics results, which by design have multiple dimensions representing the biological layers and observed physiological

information. There are methods which work with visualization tools such as Cytoscape[38], Gephi[39], etc. where users can upload datasets, along with edge and node information, to visualize 3-dimensional graphs. This requires significant experience to correctly build and visualize networks, including the steps of processing the datasets into feature coordinates suitable for the visualization tools.

Another example of a visualization method which can cover elements from multiple datasets and their relationships is a Circos plot[40]. It does an efficient job of generating an abstract view of relationships, however too many interconnections can distort the visualization, rendering it unusable. In recent years, some specialized biological databases have integrated various sources to annotate specific biomolecule interactions, such as STRING[19], [20] and STITCH[41], [42], for protein-protein and protein-chemical interactions respectively, display their interaction information in a pseudo 3D network visualization, with nodes being either proteins or chemicals and different categories of connections according to the score assigned to their connection. This serves as an inspiration for a proposed development of a multilayer PCA/correlation plot based 3-dimensional interactive plot based on a simple RGL based visualization method called Grimons[43]. Built upon the database backend of a KBCommons[44]–[46] framework, the interactive visualization will have multiple interactive elements to highlight relationships along with annotations from different layers.

Chapter 2: Maize nodal root samples collection and omics dataset generation

2.1 RNA-Seq Sample Collection and Sequencing

Maize nodal root tip samples (node 2) were collected from plants grown in two experiments – in the field (B73 and FR697) or in controlled-environment growth chambers (FR697). Samples collected from both experiments were sectioned into three regions (Figure 1) at the following distances from the root apex: Region A, 0-3.5 mm (including the root cap); Region B, 3.5-6.5 mm; Region C, 6.5-10 mm.

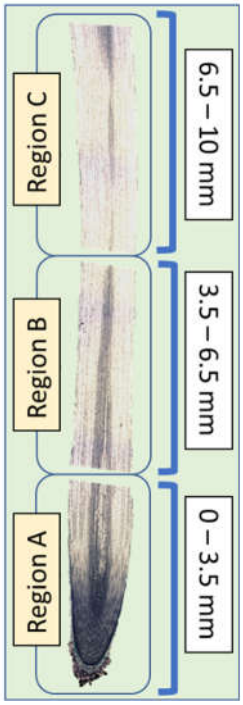


Figure 1: Maize nodal root tip structure, divided into three sections: Region A, 0-3.5 mm; Region B, 3.5-6.5 mm; Region C, 6.5-10.5 mm

Field experiments were performed at the Bradford Research Center, University of Missouri, Columbia, MO in 2017 using a protocol developed and implemented by the Sharp lab. This consisted of B73 and FR697 seeds planted at 12 seeds/m in 4.57 m plots, 4 rows wide with 0.76 m row spacing in a randomized complete block design with six replications. Plants were grown to Vegetative-stage 3, which was 16 days after planting, equating to 33.2 growing degree days (GDD) [47] after which they were harvested and root tips samples collected. The experiment was conducted under a rainout shelter, which allowed control over water availability by excluding precipitation. Well-watered plots were

irrigated at regular intervals while water-stressed plots received no water after germination. Campbell Scientific 229-L Soil Matric Potential sensors were placed at 5-15 cm into the soil in four replications to monitor soil matric potential and temperature throughout the experiment. Relevant weather data was collected from the Bradford Research Center weather station.

Growth chamber experiments were conducted in a split root system[9], [17] that was also developed by the Sharp lab at the University of Missouri. This system consists of two concentric tubes that are used as inner and outer chambers to separate the seedling (primary and seminal) root system from the nodal root system, respectively, together with the substrate (PRO-MIX HP; Premier Tech, Québec, Canada) the roots are growing in. The substrate water potentials in each chamber were independently controlled by addition of pre-calibrated amounts of water. This system was used to sample nodal root tips from FR697 plants, with the intention of analyzing biochemical responses to two water stress levels: severe stress (-0.9 MPa outer chamber, -0.4 MPa inner chamber) and moderate stress (-0.9 MPa outer chamber, well-watered inner chamber [≤-0.1 MPa]), together with a control treatment in which the substrate in both chambers was well-watered. Samples were collected 19 days after planting and germination.

The nodal root tip sections were pooled into six biological replicates for field samples and five biological replicates for growth chamber samples. Each replicate contained a minimum of eight root sections representing a minimum of four plants. Root tips were taken from Field and lab samples if their nodal root 2 lengths were measured to be within 1 (one) Standard deviation of the mean nodal root 2 length within the batch from each treatment and genotype. Root tips were frozen in liquid nitrogen and ground using a

Qiagen/Retsch tissuelyser II bead-beater with 1/8" stainless steel beads from union process (part#0070-01). Root tip homogenate was then isolated using the RNeasy Plant Mini Kit (Qiagen). Isolated RNA was DNAse-treated with TURBO™ DNase (Thermo Fisher). DNAse-treated RNA quality was assessed using the 2100 Bioanalyzer (Agilent Technologies Inc.). RNA samples were then sent to Novogene (Sacramento, CA) for library preparation and sequencing, producing high quality paired-end 150 bp RNA-Seq libraries. In total 69 RNA-Seq datasets, 34 from field-grown samples and 45 from growth chamber-grown samples were generated as summarized in table 1.

Table 1: Distribution of RNA-Seq samples replicates per phenotype and tip region for each genotype.

			Region A	Region B	Region C
Field	B73	SS	6	6	6
		WW	6	6	5
	FR697	SS	6	6	6
		WW	6	5	5
Lab	FR697	SS	5	5	5
		MS	5	5	5
		WW	5	5	5

2.2 Metabolite samples' quantification

Metabolite samples were generated from the same set of well-watered and severe stress treatment FR697 maize root tip biological samples. These samples were processed and quantified by an outside company, Metabolon, Inc. – resulting in a dataset containing 570 quantified metabolites with three replicates per condition.

Samples were prepared using the automated MicroLab STAR® system from Hamilton Company. Several recovery standards were added prior to the first step in the extraction process for QC purposes. Extraction was performed with methanol under vigorous shaking

for 2 min (Glen Mills GenoGrinder 2000) to precipitate protein and dissociate small molecules bound to protein or trapped in the precipitated protein matrix, followed by centrifugation to recover chemically diverse metabolites. The resulting extract was divided into five fractions: two for analysis by two separate reverse phases (RP)/UPLC-MS/MS methods using positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS using negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS using negative ion mode ESI, and one reserved for backup. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. The sample extracts were stored overnight under nitrogen before preparation for analysis.

The sample extracts were dried then reconstituted in solvents compatible to each of the four methods as described. Each reconstitution solvent contains a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract is gradient-eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). A second aliquot was also analyzed using acidic positive ion conditions but is chromatographically optimized for more hydrophobic compounds. In this method, the extract is gradient eluted from a C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA, and was operated at an overall higher organic content. A third aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient-eluted from the column using methanol and water, with 6.5mM Ammonium Bicarbonate at a pH of 8. The fourth aliquot was analyzed via negative ionization following

elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH 10.8. The MS analysis alternates between MS and data-dependent MS^n scans using dynamic exclusion. The scan range varies slightly between methods, but covers approximately 70-1000 m/z. These methods produced raw peak files which were analyzed using the company's own in-house informatics protocol supported by their own LIMS software and database of mass spectral entries to identify the concentration of biochemicals, identifiable as metabolites. Peaks are quantified as area-under-the-curve detector ion counts. For studies spanning multiple days, a data adjustment step is performed to correct block variation resulting from instrument inter-day tuning differences, while preserving intra-day variance. Following median scaling, then imputation of missing values, if any, with the minimum observed value for each compound, the data were transformed to the natural log for statistical analysis. The raw area count and normalized imputed datasets were reported for the 3 replicates per condition.

2.3 Representative phenotypic observations for FR697 Samples

Various measurements were also collected along with the FR697 plant tissue samples at the end of the growth cycle of the maize plants to serve as phenotypic observations. Phenotypic records taken from the plant samples consisted of multiple measurements – *Total Area, Total Mature Area, Total Immature Area, N1_Avg_Length (Nodal Root 1 Average Length), N2_Avg_Length (Nodal Root 2 Average Length), Total Plant Mass (mg), Total Shoot Mass (mg), Total Root Mass (mg), Nodal Root Mass (mg) and Seedling Root Mass (mg)*. These records were collected from a total of 294 severe stressed and 308 well-watered plants.

Using the processing steps as described in Figure 1, the phenotypic measurements were reduced to a representative set of phenotypic observations to be used as part of the multiomics data integration study. However not all plants had their measurements recorded in full and as such not all observations are statistically useful. We started by removing the

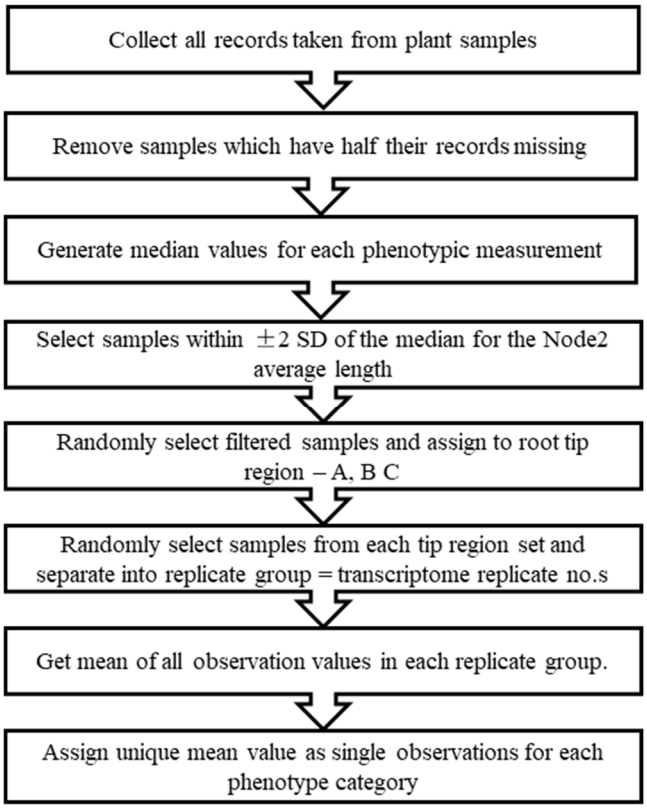


Figure 2: Workflow to reduce plant-wise phenotypic recorded measurements to a representative set of observations.

plant individuals which had records of less than half of their observations. Concurrent filtering steps retained samples within ±2 standard deviations of the median value of measured *Nodal root 2 lengths*. At the end of the workflow (Figure 2), we amassed a set of phenotypic observations which correspond to biological replicates from other omics datasets and can be used for downstream multiomics analysis. We collected measurements for an average of 12 plants per replicate for the severe stress treatment and average of 8

plants per well-watered replicate, organized into 3 replicates (Figure 3). The rationale behind this workflow is that biomatter from multiple plant samples were combined to create biological replicates for transcriptomics RNA sequencing and metabolite quantification.

No. of plants for each treatment	
SS	294
WW	308

Filtered plant samples		Rep 1	Rep 2	Rep 3
SS	A	13	12	12
	B	12	12	12
	C	12	12	12
WW	A	8	8	8
	B	7	8	8
	C	8	7	8

Figure 3: Phenotypic records and measurements of all plant samples were filtered down to average 12 plants per replicate for the severe stress treatment and average 8 plants for the well-watered control. The final representative set of observations had values which were the mean of the 12 or 8 plants per replicate.

2.4 Iso-Seq transcript dataset generation

A maize FR697 tissue collection was created using samples taken from various sections of plants grown under greenhouse conditions. These samples comprised: unpollinated silks, immature tassel, immature ear, kernels 14 days after pollination (DAP), kernels 21 DAP, whole germinated kernels, whole seedling at the 2-leaf stage, young leaf, ligule, mature leaf-base, mature leaf-mid section, mature leaf-tip, sheath, nodal root minus tip, and nodal root tip. RNA was extracted from these tissue samples using the RNeasy (Qiagen, Hilden, Germany) kit with RLC buffer following the manufacturer's recommended protocol. The RNA samples were then pooled for subsequent amplification, from which Barcoded SMRT libraries were prepared and sequenced on the PacBio platform with X SMRT cells by Novogene Corporation Inc. (Sacramento CA).

Resultant PacBio Iso-Seq reads were processed using the Iso-Seq 3 analysis pipeline (Pacific Biosciences) [48]. This included Circular Consensus Sequence (CCS) generation, full-length reads identification ("classify" step), clustering isoforms ("cluster" step), and a "polishing" step using Arrow consensus algorithm. At the end of this pipeline, a set of long read transcripts was generated to be used for further analysis.

Chapter 3: Gene expression analysis of maize drought stressed RNA-Seq samples

Analyzing the gene expression patterns for a set of samples is a robust method to gain insights into the biological properties of an organism. Gene expression analysis can broadly be broken down into two types of transcript quantification – Relative and absolute. Absolute quantification looks at transcript abundance per gene across a genome for each sample of a certain phenotype. Relative quantification compares the expression levels between groups of samples for different phenotypes to predict which genes might have significantly active in each group. Significantly expressed genes are organized into highly correlated clusters based on their expression values which are linked to annotations terms related to drought, suggesting some involvement in the maize plant's unique adaptation. The same expression datasets are used in combination with a list of known transcription factors to predict gene regulatory networks, by connecting significant transcription factors and genes.

3.1 Quantifying gene expression for a set of samples

Gene expression is a blanket term which covers the process by which information from a gene is used in the synthesis of a functional gene product through the process of translation and transcription (central dogma of biology)[49]. This results in various products such as proteins, housekeeping RNA, small non-coding RNAs (miRNAs, piRNA) and long non-coding RNA. Proteins form the bulk of these products and many of them acting as

regulatory factors. Analyzing gene expression usually involves quantifying the direct product of transcription, i.e., transcripts. Various methods have been developed over the years ranging analyzing a single or a small group of genes using northern and western blots, fluorescent in situ hybridization, quantitative PCR, etc (Low/mid plex methods) to large datasets which cover a majority of genes present in a genome. This includes DNA microarrays[50], [51], serial analysis of gene expression (SAGE)[52], [53] and more recently using RNA-Seq libraries generated from cDNA sequences[54], [55].

Absolute quantification of transcript abundance is conducted by interpolating the PCR amplification signal for a cDNA/RNA sequences onto a standard curve, and is usually presented as either a weight quantity, concentration, or the most used format – "copies" of the sequence amplified. This sort of quantification is generally used by quantitative PCR or real time PCR methods.

For high-throughput absolute quantification, short read segments of transcripts are collected from tissue bio samples and amplified via PCR. Using dedicated high-performance tools, these amplified sequences are aligned against reference gene sequences, and the expression levels are reported in terms of absolute numbers of these short reads per gene. Relative quantification involves comparing two or more sets of samples to contrast and highlight genes which have significant changes in their expression levels. Also called differential expression, results are usually reported in the form of observed log fold change in expression against the null hypothesis of no change[56]. According to the fold change, a gene is considered "upregulated" if its expression is higher in the treated samples vs. control samples; and "downregulated" if expression is higher in the control vs. treated samples.

3.2 Gene expression analysis for maize root RNA-Seq samples

To quantify transcript abundance levels, the first step in the informatics pipeline is to process the raw fastq files generated from RNA sequencing. For this the Fastqc[57] tool is used to conduct a preliminary check and report any inconsistencies in the sequencing files. Erroneous adaptor sequences, and low-quality edges are removed from reads using trimmomatic[58]. The processed fastq files are then aligned to the B73 v4 maize reference genome[59] using the HISAT2[60] read aligner using its paired end alignment function. The aligned data is analyzed using Cufflinks[61] which measures the gene expression in Fragments Per Kilobase of transcript per Million (FPKM).

RNA-Seq alignment rates for B73 field, FR797 field and FR697 lab were reported as 89.29%, 80.43% and 80.97% respectively, for an average of 84.68% mapping rate.

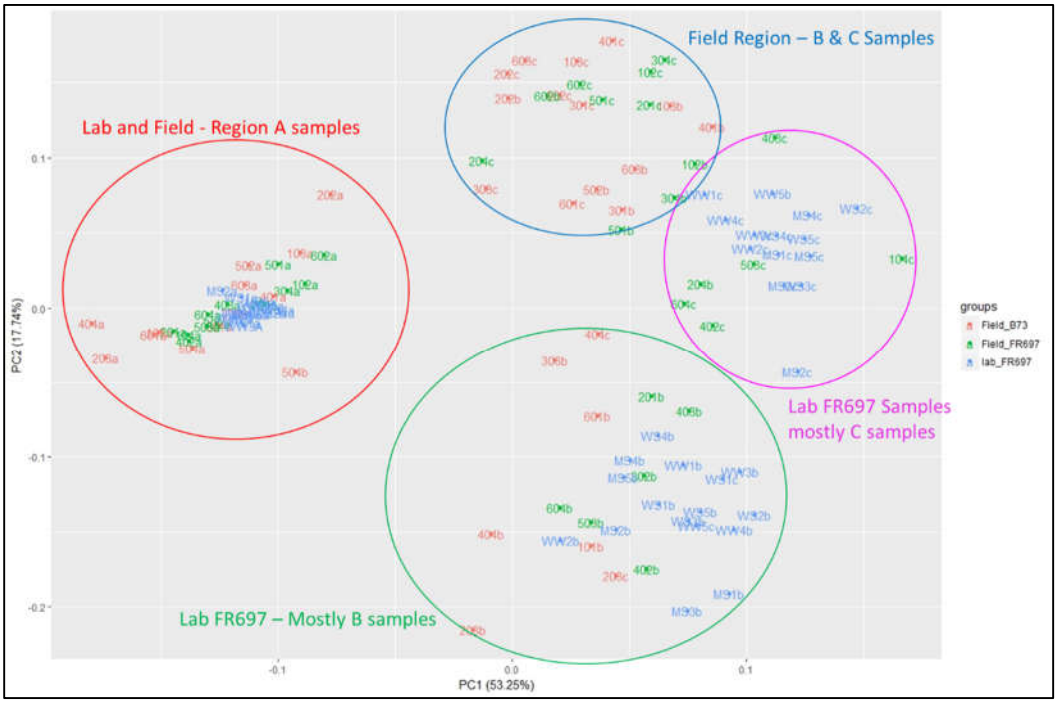


Figure 4: PCA plot for replicates' FPKM values.

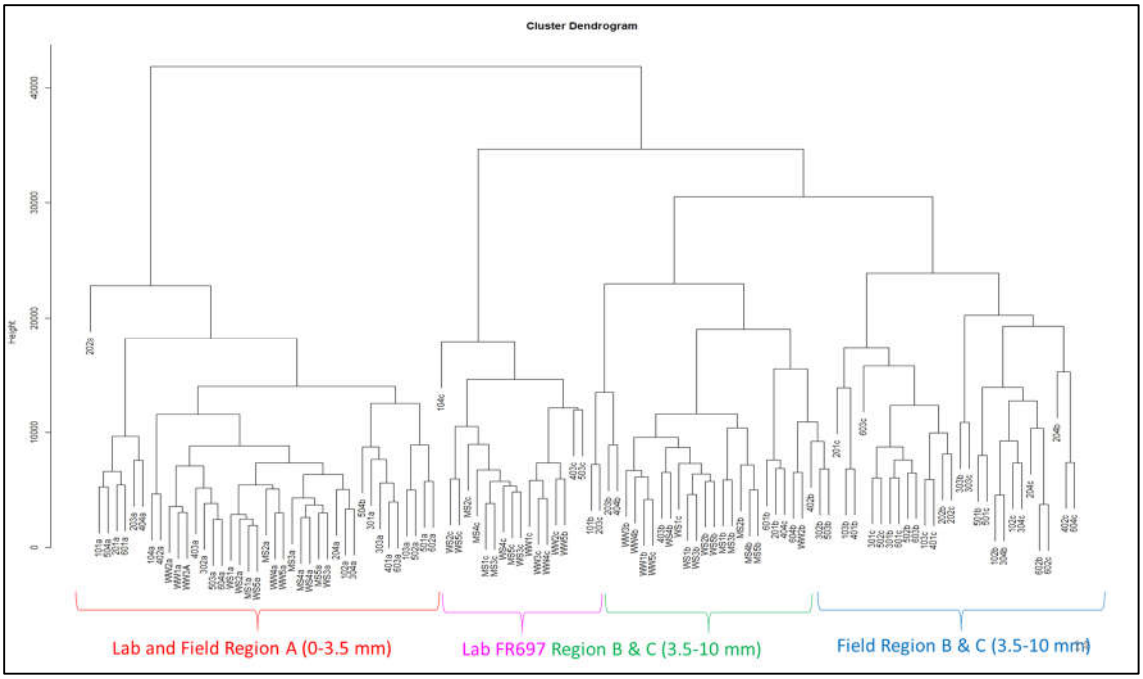


Figure 5: Hierarchical Clustering tree for replicate FPKM values.

The aligned replicates were then quantified by the first part of our RNA-Seq pipeline by Cufflinks, which is a normalized method to calculate gene expression (FPKM) & htseq-count, which is takes aligned reads to report gene expression (in terms of aligned read counts per gene). A PCA plot (Figure 4) and a hierarchical cluster dendrogram (Figure 5) show that the replicates clustered into groups, separating based on which region of the root tip the samples come from. Tip region A samples formed a separate cluster itself in both plots. Region B and C formed separate but slightly overlapping clusters, suggesting that many genes have similar expression levels, but there is still a small but significant group of genes expressed differently in each tip region.

3.3 Differential expression analysis

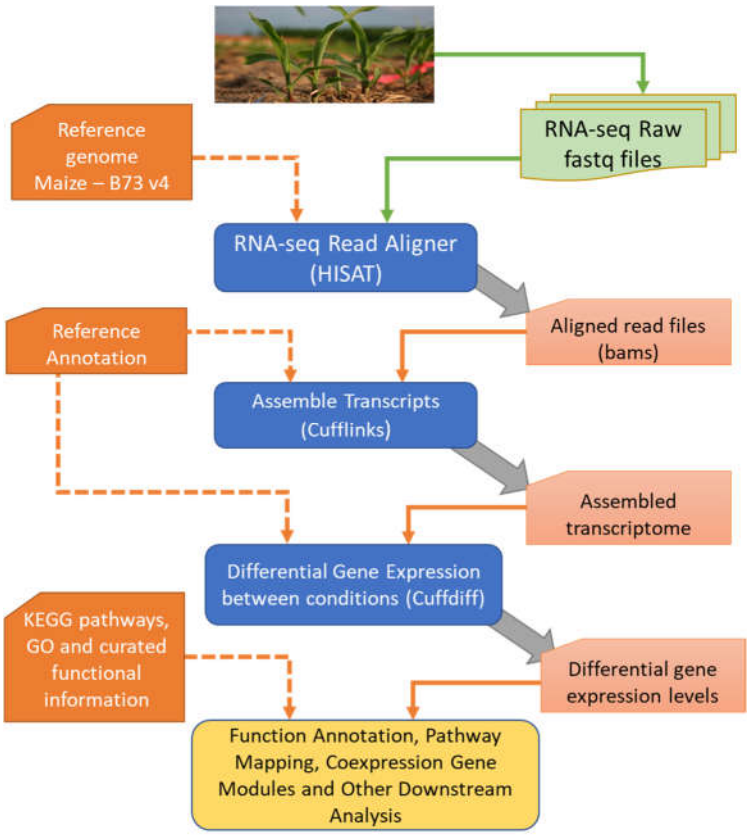


Figure 6: Transcriptome analysis and Differential expression analysis pipeline.

The differential expression part of the pipeline includes the cuffdiff method from the Tuxedo suite of tools; used to find transcript expression differences between multiple comparisons. It measures observed log fold change in its expression against the null hypothesis of no change[61]. This analysis was performed for multiple combinations – between stress levels, tip regions, and across genotypes as well. Cuffdiff already accounts for variation between replicates and batch effects, and as such no major normalization is required. Each comparison reported significant groups of up and down regulated genes, with some significant genes being conserved across the comparisons. A set of results are show in Table 2, 3& 4. Differentially expressed genes were considered significant if above a value of log fold change = 2 in either direction and p-value = 0.05. An observation was

made that several genes which code for transcription factors were shown to be differentially expressed, however were just below the fold value threshold of fold change of 2. The pipeline was updated so that such genes are retained in the results if their fold change is above 1.2 with the rationale that such gene are known to have lower than expected expression levels.

Table 2: Numbers for significantly up and down regulated genes from field samples.

FIELD							
cuffdiff_1	up	down	total	cuffdiff_4	up	down	total
SSa_vs_WWa_B73	1776	246	2022	SSa_vs_WWa_FR697	830	401	1231
SSb_vs_WWb_B73	2367	3235	5602	SSb_vs_WWb_FR697	1554	1721	3275
SSc_vs_WWc_B73	1189	2194	3383	SSc_vs_WWc_FR697	1121	2381	3502
cuffdiff_2				cuffdiff_5			
WWa_vs_WWb_B73	3119	4888	8007	WWa_vs_WWb_FR697	4811	5791	10602
WWa_vs_WWc_B73	4454	6017	10471	WWa_vs_WWc_FR697	4941	6594	11535
WWb_vs_WWc_B73	141	130	271	WWb_vs_WWc_FR697	413	416	829
cuffdiff_3				cuffdiff_6			
SSa_vs_SSb_B73	5187	5921	11108	SSa_vs_SSb_FR697	3706	5732	9438
SSa_vs_SSc_B73	4253	5341	9594	SSa_vs_SSc_FR697	4046	6180	10226
SSb_vs_SSc_B73	71	61	132	SSb_vs_SSc_FR697	129	30	159

Table 3: Numbers for significantly up and down regulated genes from lab samples.

LAB							
cuffdiff_7	up	down	total	cuffdiff_10	up	down	total
MSa_vs_WWa_lab_FR697	709	498	1207	WWa_vs_WWb_lab_FR697	5141	6434	11575
MSb_vs_WWb_lab_FR697	1188	2253	3441	WWa_vs_WWc_lab_FR697	5506	6840	12346
MSc_vs_WWc_lab_FR697	773	1453	2226	WWb_vs_WWc_lab_FR697	116	42	158
cuffdiff_8				cuffdiff_11			
SSa_vs_WWa_lab_FR697	602	759	1361	MSa_vs_MSb_lab_FR697	5586	5385	10971
SSb_vs_WWb_lab_FR697	785	1625	2410	MSa_vs_MSc_lab_FR697	5820	6854	12674
SSc_vs_WWc_lab_FR697	914	1177	2091	MSb_vs_MSc_lab_FR697	3407	3613	7020
cuffdiff_9				cuffdiff_12			
SSa_vs_MSa_lab_FR697	9	32	41	SSa_vs_SSb_lab_FR697	5642	6108	11750
SSb_vs_MSb_lab_FR697	381	400	781	SSa_vs_SSc_lab_FR697	5680	6873	12553
SSc_vs_MSc_lab_FR697	332	45	377	SSb_vs_SSc_lab_FR697	1925	2246	4171

Table 4: Numbers for significantly up and down regulated genes for a set of cross genotype comparisons.

B73_Field_vs_FR697_Field							
cuffdiff_13	up	down	total	cuffdiff_14	up	down	total
WWa_FR697_vs_WWa_B73_field	919	1299	2218	SSa_FR697_vs_SSa_B73_field	1003	2522	3525
WWb_FR697_vs_WWb_B73_field	431	975	1406	SSb_FR697_vs_SSb_B73_field	1662	2753	4415
WWc_FR697_vs_WWc_B73_field	472	323	795	SSc_FR697_vs_SSc_B73_field	1809	2449	4258

3.4 Generating gene correlation clusters and annotation

To find groups of genes with conserved expression levels over specific comparisons, we use Weighted gene correlation network analysis (WGCNA)[18]. It takes advantage of a graph theoretical approach to measure correlations amongst genes and then groups genes into modules which usually are associated with coordinated or related biological functions and regulatory mechanisms. The working of WGCNA can be summarized as:

- Builds Weighted undirected gene networks with interconnected nodes
- Nodes correspond to Gene expression profiles of significant genes – usually a list of genes from a differential expression analysis, like from the previous section
- Relationship between nodes defined by pairwise correlation between profiles
- Defined by a matrix $X = [x_{il}]$
- $a_{ij} = |\text{cor}(x_i, x_j)|^{\beta}$ ($\beta \geq 1$ (soft threshold)) – Gives priority to highly correlated pairs

From this a Hierarchical Tree is generated, and a “tree cutting” value is specified, which defines the clusters formed by the method.

WGCNA is a classic dimension reduction method, as it takes expression values (FPKM or counts) for constrained list of differentially expressed genes from the various comparisons

and conducts a pairwise comparison to build clusters (referred to as modules as well) of co-expressed genes. This will essentially reduce a large differentially expressed gene list to smaller modules of co-expressed genes which when annotated, showing significant overlap with development terms related to root tip growth. This is important since it is hypothesized that a subset of expressed genes is conserved in the drought stressed samples, especially when comparing FR697 field and lab samples, and DE genes which are grouped into significant modules, can probably be linked to responses to other environmental factors, as candidates which can be filtered out.

The input for WGCNA is a matrix of gene expression values filtered according to a combined DE gene lists from all three root tip region wherein duplicate names are removed. This aligns with the fact that while cell division occurs primarily in the A region, cell elongation occurs in all the three regions, and as such an interest is there to see if any unique clusters form with certain genes active over all the root tip regions.

Modules are assigned a color value for easy identification. GO enrichment and KEGG pathway enrichment is conducted for each module, using the AgriGO v2[62] online tool and KEGGrest[63] function from Bioconductor respectively. An example of such GO annotation for all modules is shown in table 5 & figure 7 for the **combined input from Severe Stressed vs Well Watered samples for Lab FR697**; along with KEGG pathway annotation for two specific modules from this comparison.

Table 5: Significant Ontology terms associated with each WGCNA color module for SS vs WW FR697 lab samples. **Blue, Green & Grey modules had "Response to stress" amongst the top 5 GO terms.**

SS_vs_WW_FR697_Lab		
Merged Colors	Freq	Significant Ontology
black	56	None
blue	542	response to stimulus
brown	506	DNA replication
green	354	response to stimulus
grey	29	response to stimulus
red	203	response to oxidative stress
turquoise	1894	regulation of macromolecule biosynthetic process
yellow	459	response to chemical stimulus

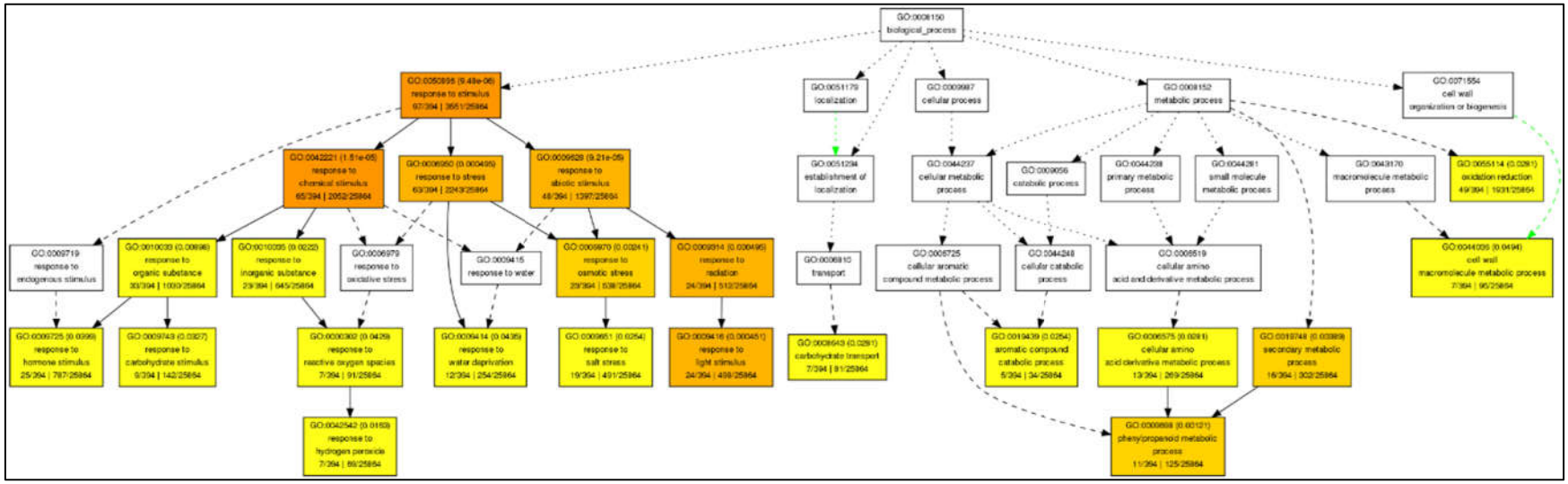


Figure 7:GO ontology hierarchical tree generated by AgriGO, Higher the number of genes with specific terms show in the boxes, the darker the color. **Blue, Green & Grey modules had "Response to stress" among top 5 GO terms**. An informative way to visualize GO enrichment for differential expressed genes.

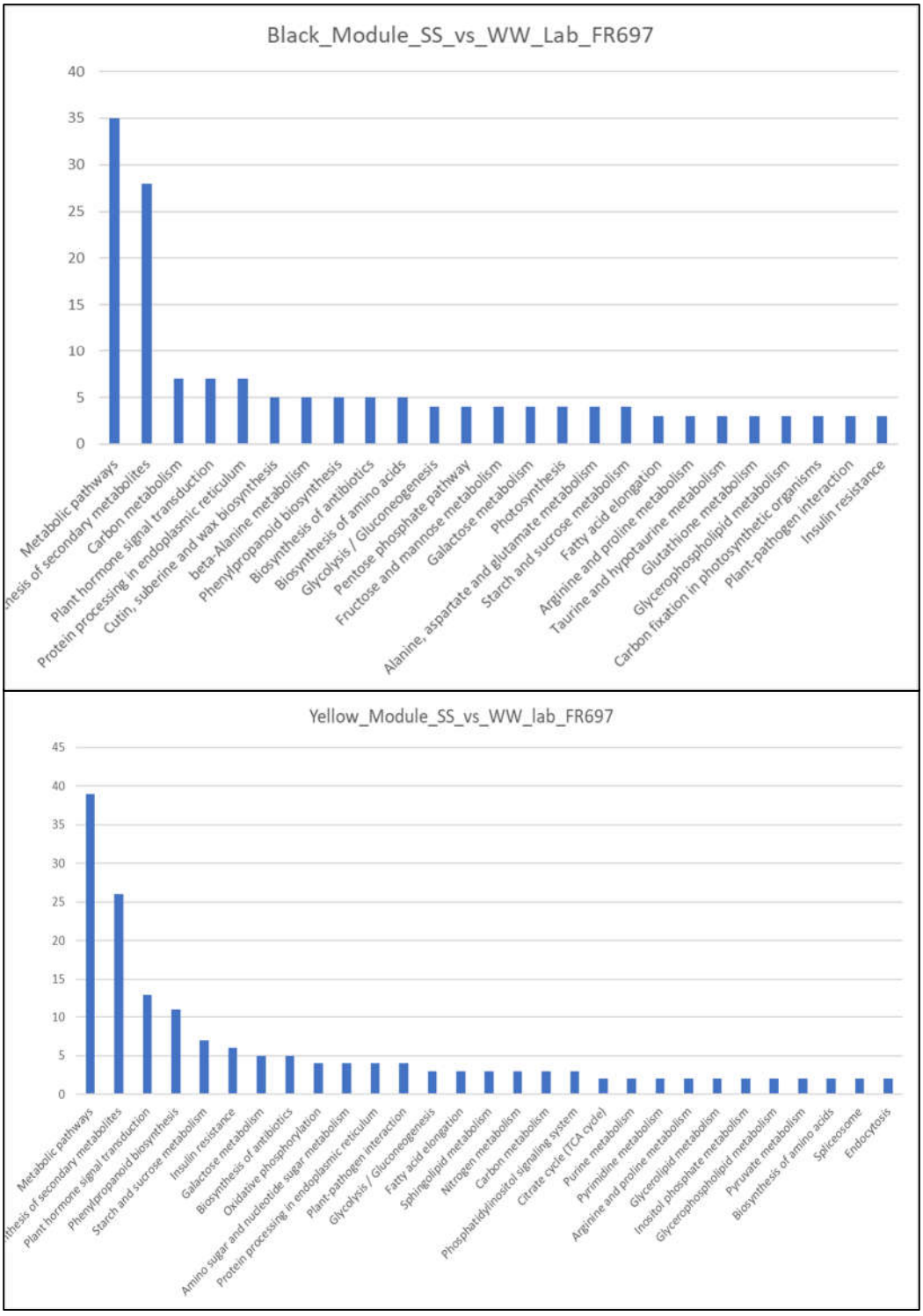


Figure 8: Two bar charts showing the distribution of WGCNA modules (clusters) of DE genes assigned to specific KEGG pathways.

3.5 Predicting Gene regulatory networks

Another level of annotation is added by predicting gene regulatory networks. This involves finding correlations between various expressed transcription factors and genes, clustered to generate networks of TF-gene interactions. GENIE3[64], another unsupervised method is used to produce a directed graph of interactions between transcription factors and genes of interest, giving insights into which transcription factors might be controlling certain gene expression patterns. The main features and working of GENIE3 are:

- Input is defined by a matrix $X = [x_{il}]$ of expression values for genes.
- The matrix comprised of expression levels for differentially expressed genes with 2-fold log change and Q-value = 0.05
- A list of genes differentially expressed with 1.2-fold log change and coding for transcription factors are used as seeds.
- A Random forest Method implementation (Regression Trees) predicts target of these Seed TFs.
- Generates a ranked list based upon the strength of putative regulatory links between a target gene and the expression pattern of TFs.
- i.e – TF_1 → Gene_1, TF_1 → Gene_2, TF_2 → Gene_1… TF_n → Gene_n

The top 1000 or 2000 interactions from the ranked list are considered significant, with lower ranked predicted interactions usually have very low support. This results in a small set of high confidence elements to form a directed network of possible TF to gene interactions. This network is visualized in Cytoscape, as described in the next section.

3.6 Visualizing the combined network in Cytoscape

Cytoscape[65] is visualization tool for biomolecular networks. It does not contain any data of its own, instead provides a unified framework where various datasets of biomolecular interactions can be loaded onto and using a set of unique identifiers common across these datasets, overlay the information on top of each. These integrated networks can be queried, and interactions can be highlighted to showcase interesting connections. The Cytoscape application is also extended by various plugins, which allow it to directly query various databases and automatically annotate networks, along with network specific visualization colors and themes.

A filtered maize protein – protein interaction network obtained from STITCHdb is first loaded in the Cytoscape application. Both clustering results – from WGCNA and GENIE3 are then added as annotation layers on this network, along with a list of transcription factors. Genes are color coded according to their WGCNA clusters and transcription factors are assigned a different shape (triangle). The connections between TF and genes are changed to directional as well. Users can select specific interactions along with predicted enriched annotations and highlight the genes with their expression levels. This allows the interrogation of the large dataset in a quick and informative manner, highlighting unique clusters if they form. An example for an interesting cluster formed around the gene Zm00001d038746: Heat stress transcription factor A-4a, is shown in the figure 9 along with table 6 showing the rank of the transcription according to the number of connections, along with table 7 showing genes interacting with this example. It reports that our TF of interest interacts with 4 expansin protein encoding genes. Expansin is found in plant cell

walls and has important functions in cell growth and is a good candidate to explore if influenced by drought stress.

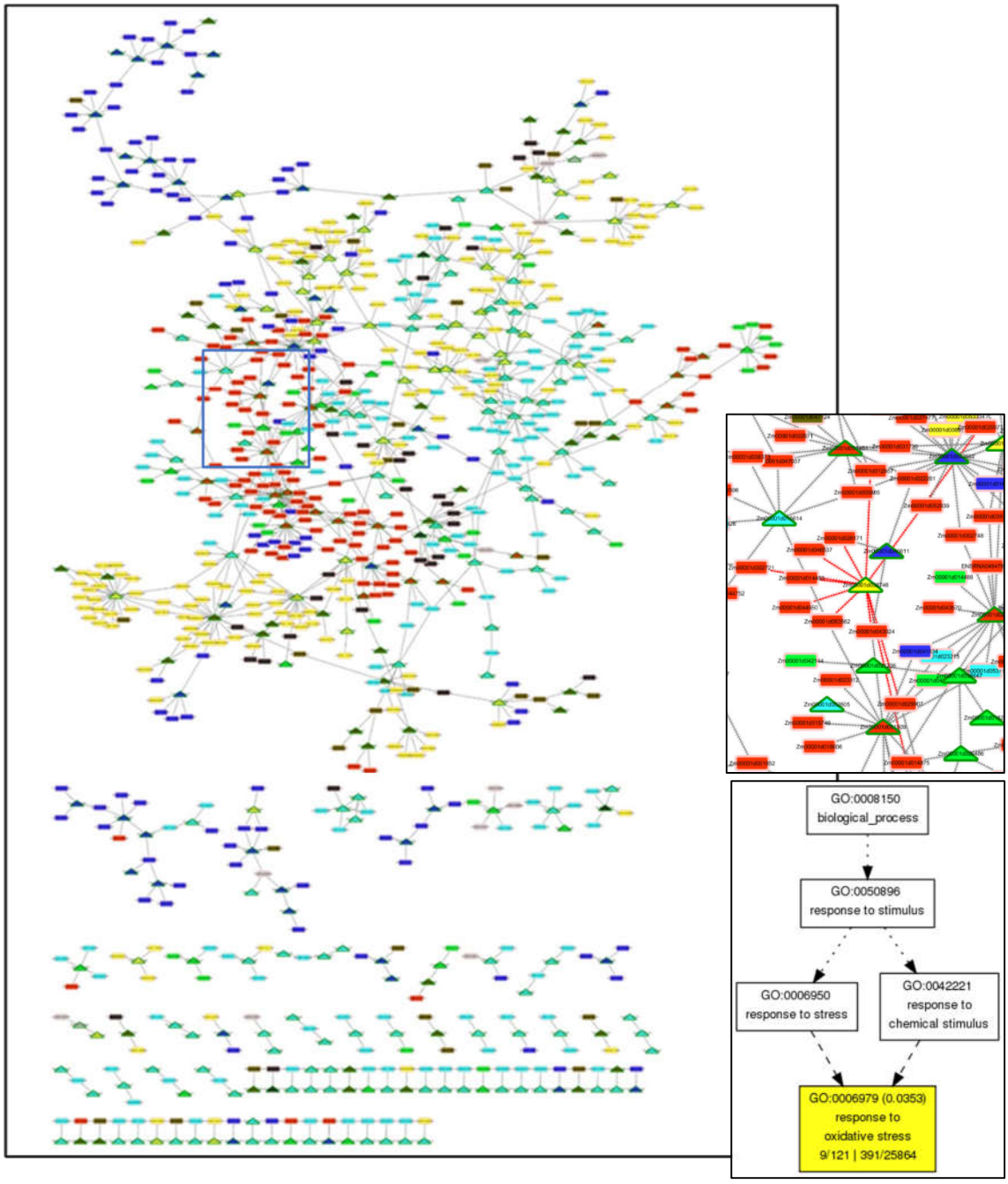


Figure 9: Network visualized in cytoscape with details from Genie3 and WGCNA for SS_vs_WW_Lab_FR697 comparison. One interesting cluster is highlighted and zoomed in the insert. Shows transcription factor coding gene Zm00001d038746 interaction with a set of genes mostly from the red WGCNA cluster. The second insert shows that most of the genes in the shown cluster are annotated with GO terms related to "response to stress".

Table 6: Table showing a ranked list of TFs with gene interactions shown in the cytoscape network for SS_vs_WW_Lab_FR697. Highlighted is TF:Zm00001d038746, a highly ranked transcription factor found interacting with a number of genes annotated for stress response.

TF	genes connected	Annotation by MaizeGDB
Zm00001d048352	23	IQ-domain 20
Zm00001d033602	21	Nuclear transcription factor Y subunit A-9
Zm00001d039510	21	IQ-domain 25
Zm00001d052087	21	Ethylene-responsive transcription factor RAP2-2
Zm00001d002630	17	Calmodulin binding protein
Zm00001d022613	17	Delayed flowering1
Zm00001d030532	15	Homeobox-leucine zipper protein ATHB-6
Zm00001d051509	15	Putative HLH DNA-binding domain superfamily protein
Zm00001d030232	14	Transcription factor bHLH18
Zm00001d036418	13	dbb9; double B-box zinc finger protein9:
Zm00001d015521	12	Two-component response regulator ARR12
Zm00001d021946	12	unknown
Zm00001d050018	12	ABSCISIC ACID-INSENSITIVE 5-like protein 5
Zm00001d012401	11	prh4; protein phosphatase homolog4:
Zm00001d017606	11	Transcription repressor OFP6
Zm00001d038746	11	Heat stress transcription factor A-4a
Zm00001d044940	11	Putative bZIP transcription factor superfamily protein
Zm00001d017831	10	AT-rich interactive domain-containing protein 3
Zm00001d031451	10	SBP-domain protein 5
Zm00001d038221	10	NAC domain containing protein 32
Zm00001d038397	10	nfy2;NF-YB homolog: single PCR-isolated sequence with strong homology to CCAAT-box binding protein subunit
Zm00001d050404	10	G2-like transcription factor

Table 7: This table shows the list of genes interacting with TF:Zm00001d038746 with their functional annotations. Four of them highlighted as extensin like protein coding along with which root tip region they come from.

Genes which are interacting with TF: Zm00001d038746		Tip Region
Zm00001d014493	Wound-responsive family protein	B & C
Zm00001d053562	Probable F-box protein	A, B & C

Zm00001d002721	extensin-like protein	A & B
Zm00001d044950	Ethylene-responsive transcription factor ERF014	A, B & C
Zm00001d020071	Sugar transport protein 14	A, B & C
Zm00001d046537	ABC transporter G family member 26	A, B & C
Zm00001d029907	Expansin-B4	A, B & C
Zm00001d053562	Probable F-box protein	A, B & C
Zm00001d012957	Expansin-A6	A, B & C
Zm00001d043024	Unknown	A
Zm00001d014875	ago1d; argonaute1d: Ortholog of Arabidopsis ago1	B & C
Zm00001d026171	Expansin-B4	B & C

3.7 Conclusion

This chapter detailed the rationale and use cases of the informatics tools in the data mining pipeline used to tease out important patterns from RNA-Seq datasets generated from drought stressed maize nodal root samples. All the results and data tables generated from the gene expression analysis study along with their annotations are submitted to a maize nodal root specific database in the KBCommons website. This provides users a suite of interactive options to explore the large datasets along with various informative plots and visualizations to help with interpretation. As more batches of RNA-Seq datasets are expected to be included in the study, possible changes, and modification to the pipeline is discussed in Chapter 7. This is to account for batch effects which inadvertently occur in such scenarios.

Chapter 4: Multiomics analysis for FR697 omics datasets

Genes, proteins, and compounds do not work independently in their own bubble, instead they are all interconnected and work as a large complex system. With the decrease in cost of data collection we are now able to explore the relationships of elements from multiple omics datasets and their effects on each other in a systems biology manner. By treating multiple omics datasets as one large wholistic network, multivariate multiomics integration methods can find unique and significant biomarkers which discriminate between various phenotypes under observation [66], [67]. These elements, when annotated and presented with accompanying predicted or previously analyzed regulatory information, give insights into the molecular mechanisms behind such observed phenotypes.

This chapter presents the development and implementation of a multiomics data integration strategy to find unique biomarkers associated with drought stress for maize nodal root tips. The pipeline includes a framework which incorporates a duo of multivariate methods – sparse Generalized Canonical Correlation Analysis (sGCCA)[68] and generalized Partial Least Square Discriminant Analysis (PLS-DA)[69]. The method is on a sample dataset of phenotypic measurements, gene expression levels and metabolite quantification, with the further intent to integrate protein expression levels as they become available in future.

4.1 Multivariate methods for multiomics data integration

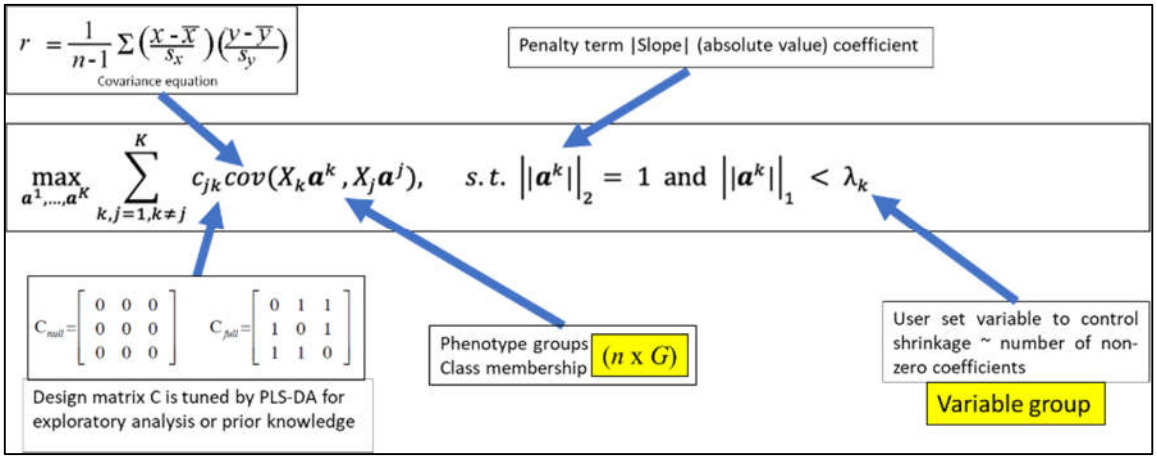
Multivariate methods, both supervised and unsupervised have been used to tease out significant elements from large, interconnected omics datasets[70]–[75]. These methods use a set of datasets that are heterogenous but generated from the same biological samples.

These elements have the potential to generate new connections between pathways and groups to serve as biomarkers related to specific phenotypic change, also called “features” in the context of statistical algorithms. Use of support vector machine-based algorithms, partial least square discriminant analysis and Random forest methods are popular feature selection methods being employed. The advantage of such algorithms is the inclusion of multiple independent variables and the influence they exert on dependent variables. However, large high-quality datasets are required to accurately predict the relationships between the variables, but this also comes with the risk of overfitting. To overcome this drawback frameworks incorporating multiple algorithms reduce large datasets down to select features, with the models being built over multiple steps. Frameworks like this also include tools to help disseminate the results generated. The Mixomics package[76] is one such framework, it incorporates multiple algorithms, both supervised and unsupervised to work with several datasets, with the actual number and quality (i.e. sparse or not) of the datasets determining which combination of algorithms are to be used. It is implemented in the R programming language and part of its Bioconductor suite of bioinformatics packages.

4.2 Selecting features of interest using a discriminant analysis method

The DIABLO method is a supervised N-integration framework, part of the Mixomics package. N-integration combines multiple datasets from the same individuals but with varying number of observations per set. It conducts a discriminant analysis to find clusters of significant elements spanning the multiple omics layers under study. It builds on two main regression analysis methods – Sparse Generalized Canonical Correlation Analysis (sGCCA)[77] and Partial Least Squares – Discriminant Analysis (PLS-DA)[69], [78].

Equation 1: Description of the optimization function for DIABLO method. The equation calculates the covariance between Omics datasets X, from k to j, with their associated coefficient 'a', shrinkage controlled by λ. Relationship level between omics layers and components defined by Cjk.



The method accepts $N\ X\ p_n$ datasets where N is the same set of samples, and p_n is the set of elements in each omics level (can be different number). The main feature of this method is the ability to predefine a bias in terms of L1 score penalty design matrix (also called design matrix) to simulate the level of "interconnection" between omics datasets i.e., the biological relationships between the layers. It extends sparse generalized canonical correlation analysis (sGCCA) for classification and uses PLS-DA which tunes Parameters for the Design Matrix. Discriminant analysis builds a predictive model for group membership (Equation 1). The model is composed of a discriminant function (or, for more than two groups, a set of discriminant functions) based on linear combinations of the predictor variables that provide the best discrimination between the groups. Specifically, sGCCA reduces dimensions to finds clusters of highly correlated elements on each omics layer, and PLS-DA maximizes the correlation between each of these reported clusters – finding the best possible overlap and assigning specific elements from each dataset.

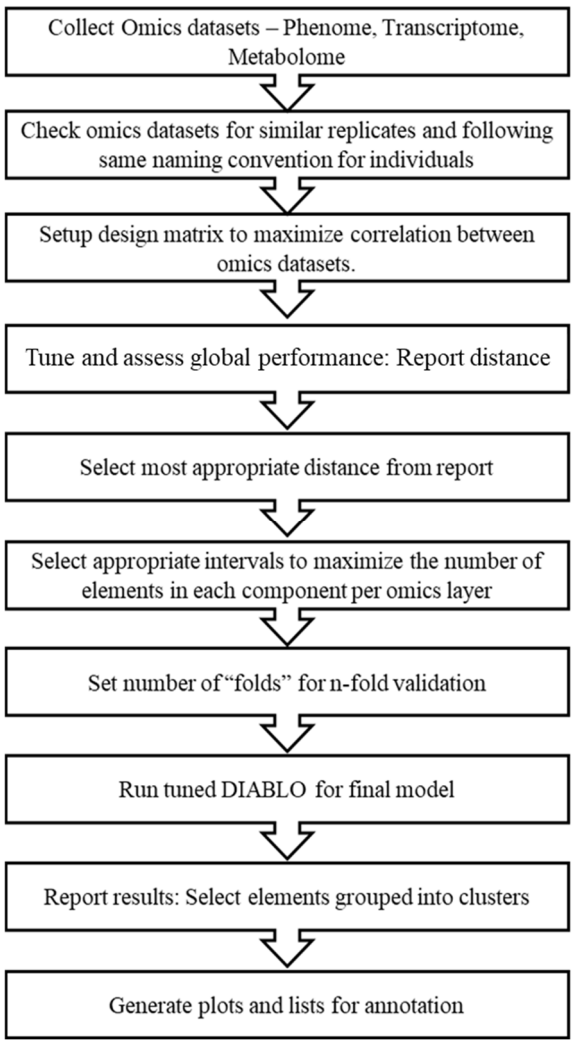


Figure 10: Workflow for using DIABLO with Plant drought root data – including tuning parameters and data compatibility checking.

Figure 10 shows the workflow applied using DIABLO as the main component in the multiomics analysis strategy, using the transcriptome, metabolite, and phenome datasets generated from the maize root samples, as detailed in chapter 2. Initial checks are performed to determine the presence of same number of individuals (i.e. same number of replicates for each treatment), and then set up a design matrix to define the relationships between the omics layers. The main function in its R method – "*sgccda.res"* is tuned by first running the function and reporting the distance, which is essential in reducing the error rate of assigning elements to clusters; and maximizing the number of clusters allowed,

before error rates increase. Then the "*keep.x"* function is used to report back a small range of numbers for elements allowed to be selected from each omics level in each cluster – such that the computation time is reasonable and viable enough to be distinct. Once the parameters were identified, "*sgccda.res"* is run again with the new parameters and set number of fold validation steps. This results in the formation of a few clusters of elements (also called components) which are candidates for biomarkers to distinguish between specific comparisons.

4.3 Test run using maize nodal root datasets

Following the steps in the workflow described in Figure 10, a sanity check was conducted on all three datasets. This was followed by assigning values of 0.1 to the design matrix, maximizing the relationship between the datasets as they were all from the same representative biological samples. The global performance was assessed and the best

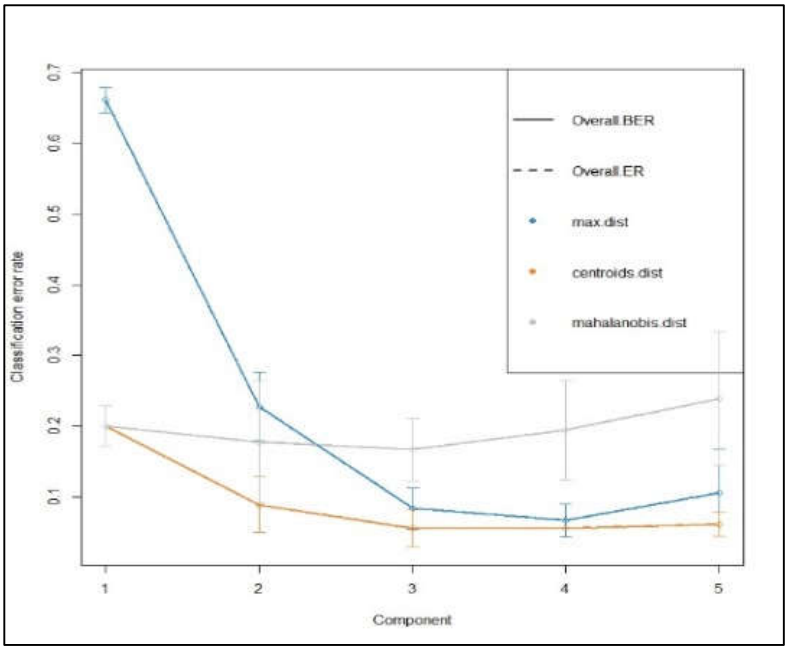


Figure 11: Plot showing all three measured distance metrics. Centroid distance has the least error rate and plateaus out at 3 components suggesting that either 2 or 3 components would capture the most significant elements.

distance metric was selected = "*centroids.dist"* and number of components set to 3 based upon the results shown in the plot in Figure 11.

The intervals for keepX and set the number of folds at 5-fold since we were dealing with small datasets. Once tuned, the DIABLO function - *tune.block.splsda* was run to perform discriminatory analysis. The method ran for ~ 72 minutes on a workstation pc with 8 cores and 16 gb ram, returning two components separating out significant elements connected to two dimensions (or observations).

Component 1 was reported as a cluster of 50 elements spread across the 3 omics layers, separating out on elements which contributed to discriminating between well-watered or severe stress samples; 26 elements were shown to positively correlate to the well-watered treatment, while remaining 24 highly correlated to the severe stress treatment. Component 2's elements separated out between Root Tip A region vs. B & C, with 47 elements highly correlated to Tip A region, and 5 elements negatively to B & C. (Figure 12). The plot shows elements correlated to specific observations in both components – specifically 6 phenotypic observations – *Total_Plant_Mass, N2_Avg_Length, Total_Root_Length, Total_Shoot_Length, Nodal_Root_Mass* and *Seedling_Root_Mass*. This indicated that these elements were influenced by elements from both clusters.

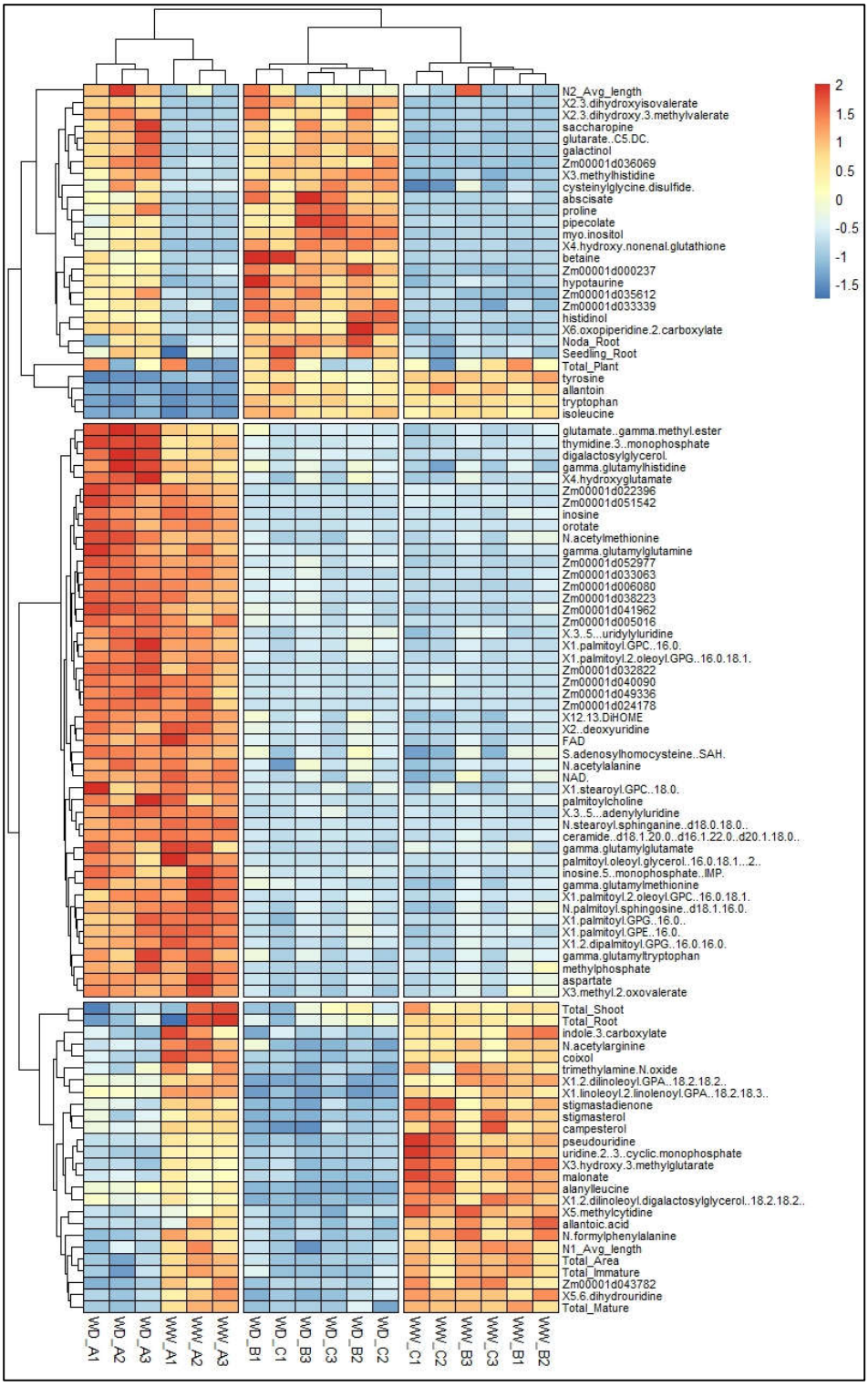


Figure 12: Cluster Plot: Showing the two components. Half of Component 1's elements showing strong positive correlation to Well-watered samples and the other half to Severe Stressed Samples. Component 2's elements all are strongly correlated to Root Tip Region A samples, and negatively to Region B & C.

The filtered correlated genes became candidates for exploration for if they are connected to any known biochemical pathways or annotated in anyway. Correlated metabolites were also connected to genic pathways and checked to determine if any were associated with specific annotations. Figure 13 exhibits a correlation circos plot for Component 1, filtered above 0.9 correlation "*r*" value. Some genes were highly correlated to *N2_Avg_Length (Nodal Root 2 Average Length),* however the stringent cutoff levels filtered out most of the results related to this observation. This was expected and considered an indication of the fact that plants struggle to grow under drought stress and do not allow for a significant number of phenotypic measurements to be collected from such plants.

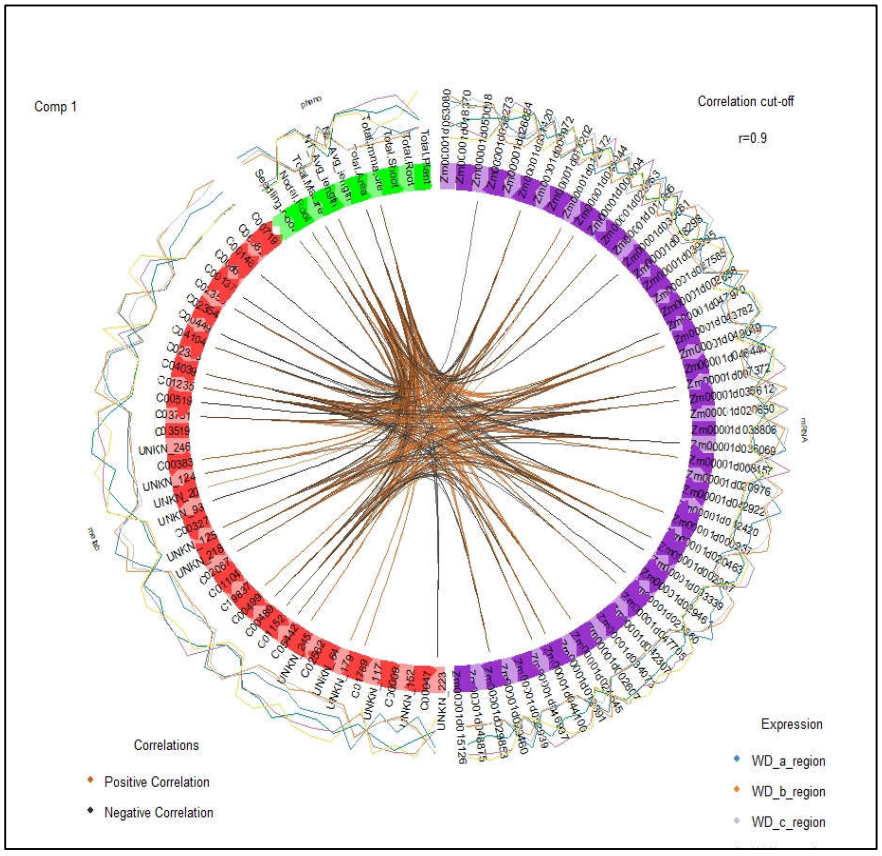


Figure 13: Circos Correlation Plot for component 1: Cutoff set at r=0.9, i.e., showing highly negative or positive correlations. Purple elements are gene ids, red elements are metabolite ids and green elements are phenotypic observations. (High resolution version: https://git.io/JkHnT)

As an alternative, the correlation matrix was reduced to elements highly correlated to *nodal_root_mass* (Table 8), with the rationale that if any significant number of root samples were collected from a plant, the recorded mass would be an indicator. *N2_avg_length* also seemed to have a decent correlation ($r \sim 0.69$) to the filtered elements. Specifically, to the predicted gene *Zm00001d035612*, which is annotated in maizeGDB[79] as a putative "inositol polyphosphate phosphatase family protein", predicted to be part of the phosphatidylinositol metabolic pathway by Reactome[22]. This pathway has known effects in several plant cellular processes, including cell growth and proliferation. Another predicted gene - *Zm00001d00228*, which was negatively correlated to *N2_avg_length*, was predicted to be part of the MAPK signaling pathway, which is involved in signal transduction, suggesting that the gene may be involved in relaying drought stress signals. These two predicted genes were found to be highly expressed across the replicates, suggesting that they may play an important role in drought stress adaptation.

Table 8: Correlation matrix filtered for nodal_root_mass for component 1.

	Nodal_ Root_Mass	N2_Avg_length	Seedling_ Root_Mass	Total_ Mature_Mass	Total_ Immature_Mass	N1_Avg_length
C01152	**0.925234324**	0.697450001	0.914356784	-0.96737923	-0.962979807	-0.950749616
C04039	**0.923005958**	0.695770238	0.912154616	-0.965049361	-0.960660533	-0.948459798
C01235	**0.922877266**	0.695673229	0.912027437	-0.964914807	-0.960526591	-0.948327558
C00489	**0.911805227**	0.687327026	0.901085567	-0.95333843	-0.949002862	-0.936950183
UNKN_125	**0.903128398**	0.680786354	0.892510748	-0.944266368	-0.939972057	-0.928034073
Zm00001d035612	**0.901489294**	0.679550782	0.890890913	-0.942552601	-0.938266084	-0.926349767
C04104	**0.901204322**	0.679335968	0.890609292	-0.942254649	-0.937969487	-0.926056937
C03761	**-0.902914997**	-0.68062549	-0.892299855	0.944043246	0.93974995	0.927814787
C02355	**-0.902931888**	-0.680638223	-0.892316547	0.944060906	0.93976753	0.927832143
C05442	**-0.905150662**	-0.682310755	-0.894509237	0.946380747	0.94207682	0.930112105
Zm00001d002287	**-0.906237661**	-0.683130145	-0.895583457	0.947517259	0.943208164	0.931229081
C02067	**-0.90656599**	-0.683377642	-0.895907925	0.947860544	0.943549887	0.931566464
UNKN_218	**-0.914286108**	-0.689197137	-0.903537282	0.955932318	0.951584952	0.939499481
N1_Avg_length	**-0.919579194**	-0.693187113	-0.90876814	0.961466506	0.957093973	0.944938535
Total.Immature	**-0.931408417**	-0.702104088	-0.920458291	0.973834556	0.969405775	0.957093973
Total.Mature	**-0.935663604**	-0.705311687	-0.924663453	0.97828357	0.973834556	0.961466506
Total.Area	**-0.939221748**	-0.707993848	-0.928179765	0.982003789	0.977537856	0.965122773

Component 2, as mentioned before, separated into the root tip region A vs. tip regions B & C. Only one phenotypic observation – *Total_plant_mass* was reported to be significantly correlated, when correlation measure "*r*" was dropped to 0.7 (Figure 14). A set of 6 genes were found to be negatively correlated to plant mass. *Zm00001d002546* was predicted to code for Histone H4, which is connected to transcription regulation and DNA repair. *Zm00001d011642, Zm00001d014756, Zm00001d022226, Zm00001d048299, Zm00001d052749* are all predicted proteins and no major annotation is associated with them (Table 9). Uniprot only reported predicted protein chains, assigning them the lowest annotation score:1 out of 5, i.e., only having assembled transcript evidence. This preliminary exploration of the highly correlated genes showed that that the multiomics pipeline worked as intended and returned actionable results. Genes from component 1 were very promising for candidate biomarkers.

Table 9: Correlation matrix filtered for Total_plant_mass for component 2.

COMP_ids	Total.Plant	N2_Avg_length	Total.Shoot	Nodal.Root	Seedling.Root
C00078	0.74827989	-0.105733812	-0.421284715	-0.126022848	-0.117088992
C02350	0.74482975	-0.105246299	-0.419342271	-0.125441787	-0.116549123
C00407	0.741482664	-0.104773347	-0.417457848	-0.124878082	-0.116025379
C00037	0.739749893	-0.104528503	-0.416482291	-0.124586254	-0.11575424
~	~	~	~	~	~
Zm00001d002546	-0.758667134	0.107201556	0.427132777	0.127772234	0.118714363
Zm00001d011642	-0.760481401	0.107457916	0.428154217	0.128077787	0.118998255
Zm00001d014756	-0.764180343	0.107980586	0.430236736	0.12870075	0.119577056
Zm00001d022226	-0.76309289	0.107826926	0.429624496	0.128517605	0.119406894
Zm00001d048299	-0.762634993	0.107762224	0.429366698	0.128440488	0.119335243
Zm00001d052749	-0.764868509	0.108077825	0.430624177	0.128816649	0.119684739

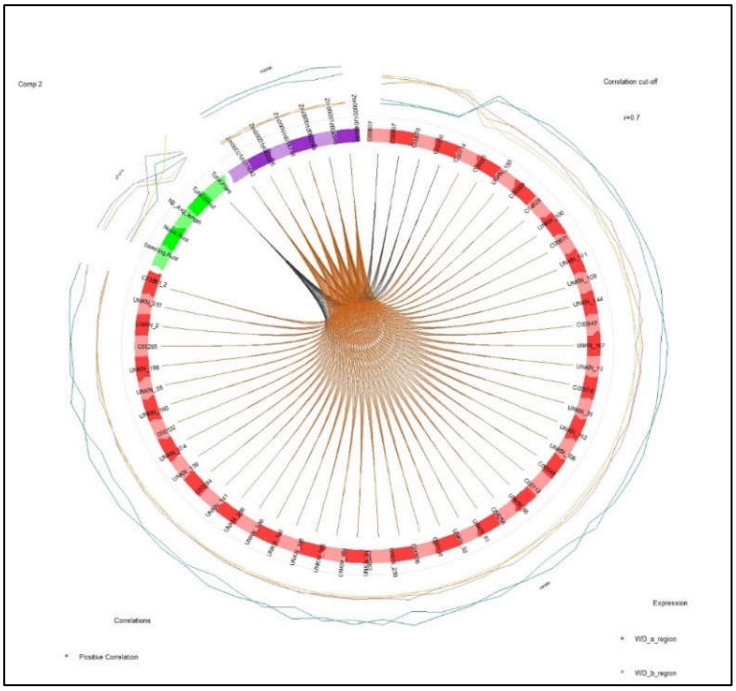


Figure 14: Circos Correlation Plot for component 2: Cutoff set at r=0.7, Only one major phenotype observation – total_plant_mass, is connected to other elements. Purple elements are gene ids, red elements are metabolite ids and green elements are phenotypic observations. (High resolution version: https://git.io/JkHnn)

4.4 Conclusions

This chapter describes the development and testing of an informatics strategy to integrate and mine the various omics datasets generated from the FR697 drought stressed maize nodal root samples. This includes using the DIABLO integration method from the Mixomics package to find clusters of significant elements. The pipeline was tested using the available FR697 gene expression, metabolite, and phenotypic observation datasets, generating significant results which correlated to annotation terms and pathways associated with cell growth and cell development. The pipeline was designed with the intent to incorporate proteomics data from the FR697 biosamples and which will be integrated when the samples are processed and ready to use. The results also served as inspiration for some of the visualization application developments detailed in chapter 6.

Chapter 5: De-novo transcriptome assembly for the FR697 genotype

As part of the effort to understand the maize plant's remarkable ability to maintain root growth under drought stress, RNA-Seq samples from drought stressed B73 and FR697 nodal root tips were previously aligned to the B73 reference genome as part of a gene expression study. Certain genes showed significantly different expression levels in FR697 compared to their B73 counterparts. However, an average of 10% difference in alignment rates was observed between the two sets of RNA-Seq reads. Some of these unaligned reads were found to be associated with proteins from the plantae proteome. It was concluded that a FR697 specific transcriptome would be useful in not only characterizing any unique transcripts or genes specific to the nodal roots but also as a future resource for drought related studies.

The processed FR697 RNA-Seq paired-end samples generated from nodal root samples as previously as described in chapter 3 were combined with Iso-Seq transcript sequences obtained from an independent set of FR697 maize tissue samples to generate a de-novo transcriptome assembly, described in chapter 2. This assembly was clustered with transcripts from a B73 reference genome guided assembly using the same set of RNA-Seq samples and the Viridiplantae proteome, to generate a consensus annotated set of "SuperTranscripts", collectively called a "SuperTranscriptome" [80]. This dataset can be used as a surrogate for a FR697 reference genome, enabling various comparative studies with the reference genotype B73 and the nested association mapping (NAM) founder lines

designed to explore the underpinning genetic networks that control nodal root growth responses under drought conditions.

5.1 Assembling a de-novo transcriptome

De-novo transcriptome assembly attempts to reconstruct a complete set of transcripts present in a dataset of reads without the aid of a reference genome.[81] Most assembly projects align together high quality RNA-Seq short read datasets to identify all possible expressed transcripts, with lengths as close to full length mRNA sequences as possible. Assembly tools where first designed with pairwise overlap alignment algorithms, however as RNA short read sequencing become more cost effective, larger datasets were used to try and capture transcripts with low levels of expression. As such the assembly tools have evolved to use greedy path[82], [83]; overlap, layout, and consensus (OLC)[84]–[86], or K-mer[87]–[89] based graph algorithms to efficiently map and build transcripts. This renders these tools computationally efficient, but results in the generation of fragments due to variation in supporting read evidence cover. Transcriptome assemblers try and overcome these by implementing a combination of these techniques.

The Trinity assembler[32] uses de Bruijn based K-mer graphs to process large number of RNA-Seq reads to build a high-quality transcriptome capturing a majority of transcripts present in the samples. Datasets are partitioned into multiple graphs to represent the transcriptional complexity for each gene or locus. Each graph is processed individually, and clusters of similar sequences are reported as unique transcripts. The size of the cluster usually represents if the reported transcript is the main isoform or not. This is reported in the unique annotation number Trinity assigns to each sequence in the output file. The

Trinity assembler has recently been updated to utilize long read sequencing evidence such as Pacbio iso-seq transcripts to resolve complex transcripts and improve on spliced isoforms, essentially filling in gaps and connecting split transcripts.

5.2 Organizing transcripts into SuperTranscripts

SuperTranscripts are an alternative representation of genes in the context of de-novo transcriptome assemblies. Redundant transcripts are collapsed together and connected using common sequence regions from spliced isoforms into a single sequence, comparable to gene's coding sequence structure. SuperTranscripts can be combined into a "SuperTranscriptome", to serve as a surrogate for a reference genome in studies that include gene expression analyses and to identify polymorphisms.

The process starts with building a directed graph for each base in a transcript's sequence. The BLAT UCSC tool's offline implementation is used to align sequences to each other, and shared bases are merged. If any forks appear after merging, they are simplified by insertion by the Lace tool between the merged/shared node bases. This usually results in extended sequences with large areas of merged bases and with small sections of inserted nodes, comparable in quality to genes.

SuperTranscriptome assembly is simplified by the Necklace pipeline[90] which combines the usage of multiple alignment and reconciliation tools to build SuperTranscripts, This includes generating a reference guided transcriptome assembly using a closely related reference genome, and finding consensus between actual genes and SuperTranscripts, supported by protein sequence evidence from a related proteome. A slightly modified

implementation of the Necklace pipeline was used to generate the FR697 de-novo SuperTranscriptome assembly, which is detailed in the next section.

5.3 SuperTranscriptome assembly pipeline

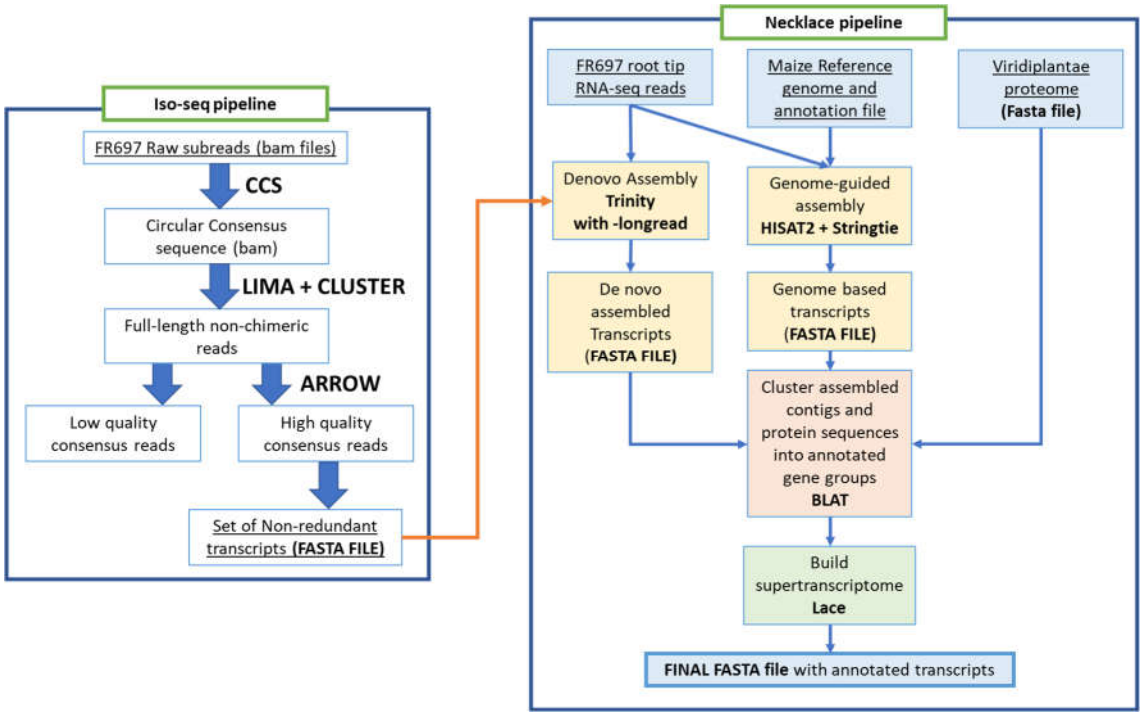


Figure 15: Overview of the steps to assemble the FR697 transcriptome. The first part consists of the Iso-Seq pipeline to generate a set of long read transcripts from Iso-Seq reads. These transcripts were used along with the RNA-Seq short read samples for a trinity transcriptome assembly. The trinity assembly along with a reference guided transcriptome assembly and the Viridiplantae proteome is combined by the Necklace pipeline, which generates a consensus "SuperTranscriptome" assembly, with the final output being a transcriptome fasta file with annotations.

FR697 SuperTranscripts were generated using the Necklace pipeline[90]. The pipeline consisted of three major steps, as follows: 1) a De-novo transcriptome assembly with Trinity (v. 2.7.0)[32] with "*longreads"* option to include the Iso-Seq transcripts; 2) a reference genome guided transcriptome assembly using the B73 v4 reference maize genome; and 3) the Viridiplantae clade proteome dataset to annotate the transcripts not

included in the provided reference gtf/gff3 files. The fully modified pipeline used is presented in Figure 3. The CD-HIT-EST program (v. 4.8.1) [33] was used for three iterations with default parameters (similarity 95%) to reduce transcript redundancy in the Trinity assembly and to compare with the results of the Necklace pipeline.

5.4 SuperTranscript annotation & verification

The intermediate trinity de-novo transcriptome assembly was generated with an ExN50 maximum value of 2173 at Ex = 95 and transcripts = 38512. The intermediate genome guided transcriptome assembly, using HISAT2, reported an alignment rate of 84.08% for the RNA-Seq reads to the maize genome. Finally, the Necklace pipeline produced 47915 unique SuperTranscripts. Of the total, 42612 were assigned unique maize reference gene IDs by the pipeline. Of the remaining 5303 SuperTranscripts, 1592 were annotated as tRNAs, 325 as a combination of mitochondrial and chloroplast genes, 3258 as novel unknown transcripts, and 128 IDs that were predicted to transcribe for proteins found in the Viridiplantae proteome, suggesting they were FR697 genotype-specific novel genes (Figure 4). The N50 of the assembled SuperTranscripts was significantly improved compared to trinity transcripts, from 1589 to 3152, which was close to the average size of maize genes of 4 Kb. The number of redundant transcripts was significantly reduced when compared to the original trinity assembly and the results of CD-HIT-EST after three iterations (Table 1).

Blastn[91], [92] was used to compare the maize gene id annotated SuperTranscripts to the actual coding region sequences of maize genes in the B73 genome. All annotated SuperTranscripts were found to be in the top three blastn hits and within 93% identity

threshold of maize genes with the same IDs. We then used HISAT2 to align a representative subset of the nodal root RNA-Seq samples against the assembled SuperTranscriptome. The alignment rate averaged at approximately 85% for all samples. This was a significant increase from an average of an 80% alignment rate when the same samples were aligned against the reference B73 v4 genome.

Table 10: Various metrics for the FR697 SuperTranscriptome assembly compared to three maize reference genotypes and the improvement of annotation over the initial trinity assembly

	MaizeMo17_CAU_scaf	MaizeB73_v4_scaff	MaizeW22_NRgene	Trinity Transcripts	Trinity with cd-hit 3 t	FR697 SuperTranscr
Number of scaffolds	2560	596	306	720299	540032	47935
Total size of scaffolds	2182615441	2134339606	2133868603	587326543	402421838	98256966
Total scaffold length as percentage of assumed ge	99.20979277	97.01543664	96.99402741	24.47193929	16.76757658	4.09404025
useful amount of scaffold sequences (>=25K nt)	2166421525	2134248774	2132523330	58570	58570	0
% of estimated genome that is useful	98.47370568	97.01130791	96.93287864	0.002440417	0.002440417	0
Longest scaffold	32176138	39317442	83688764	29921	29921	22234
Shortest scaffold	1007	5568	711	176	182	32
Number of scaffolds > 1K nt	2560	596	305	149963	97893	31091
Number of scaffolds > 10K nt	2216	591	291	510	320	271
Number of scaffolds > 100K nt	475	366	130	0	0	0
Number of scaffolds > 1M nt	304	296	97	0	0	0
Number of scaffolds > 10M nt	69	69	62	0	0	0
N50	10204498	10679169	35520101	1589	1336	3152
L50	69	62	19	95173	72937	9621
NG50	9989738	10214929	33636442	0	0	0
LG50	70	66	20	0	0	0
%A	26.16202361	26.17515251	26.11362927	24.76932496	24.77012567	24.91816509
%C	23.04310739	23.08858481	22.92158314	25.33622867	25.32836998	24.42904761
%G	23.03462569	23.10360491	22.93553555	25.01578785	24.9863945	24.9521983
%T	26.15118015	26.1942841	26.12596901	24.87865851	24.91510985	25.70053812
Total Number of Ns	35119661	30699779	40613559	0	0	50
%N	1.609063161	1.438373674	1.90328303	0	0	5.09E-05

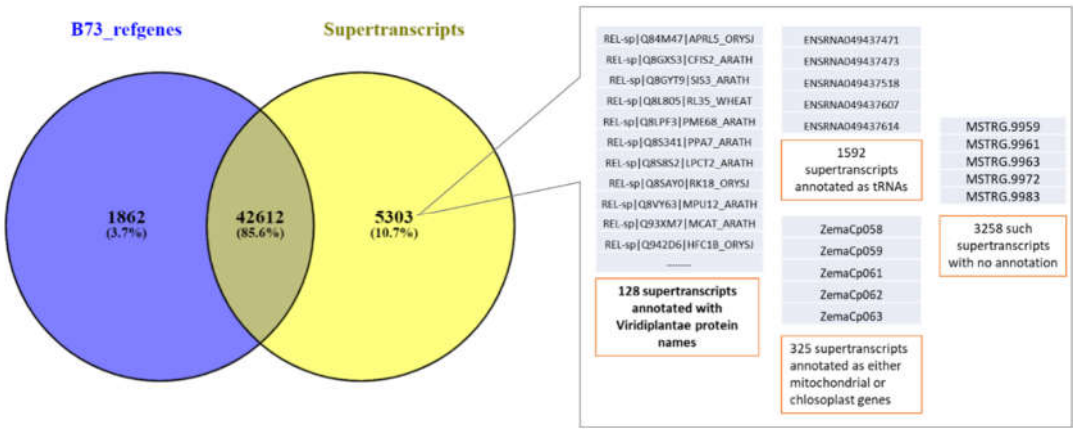


Figure 16: Venn Diagram showing that 42612 SuperTranscripts identified and annotated with corresponding maize reference gene ids. Of the remaining 5303, 128 SuperTranscripts were predicted to be coding for certain proteins from the Viridiplantae proteome, 1592 were annotated as tRNAs, 325 were annotated as either mitochondrial or chloroplast genes, and the remaining 3258 were unidentified.

5.5 GO annotation for SuperTranscripts coding for Viridiplantae proteins

To further annotate the 128 SuperTranscripts identified to code for Viridiplantae proteins, we used Blast2GO software [93] to associate GO annotations to them. The top blast hits were selected for each SuperTranscript (blast E-value very close to 0), and the corresponding distributions of annotations for biological processes, molecular function and cellular components are reported in Figure 5. The top terms from all three GO categories point to these new sequences having some connection to oxidoreductase activity in the cellular membrane. A broad literature search for this family of enzyme suggested that it is integrated with many cellular pathways[94] related to root activity, along with some studies suggesting a connection to stress responses[95], [96]. However, a deeper exploration including biological validation will be required to confirm a direct role in the drought stress response. In addition, some of the SuperTranscripts were assigned GO terms related to heme binding, which might reflect the fact that iron rich water was used to saturate the soil in chambers used in the lab method i.e., the split root system. A smaller group of SuperTranscripts were assigned terms related to biological processes related to responses to heat and cold, and stimulus to light.

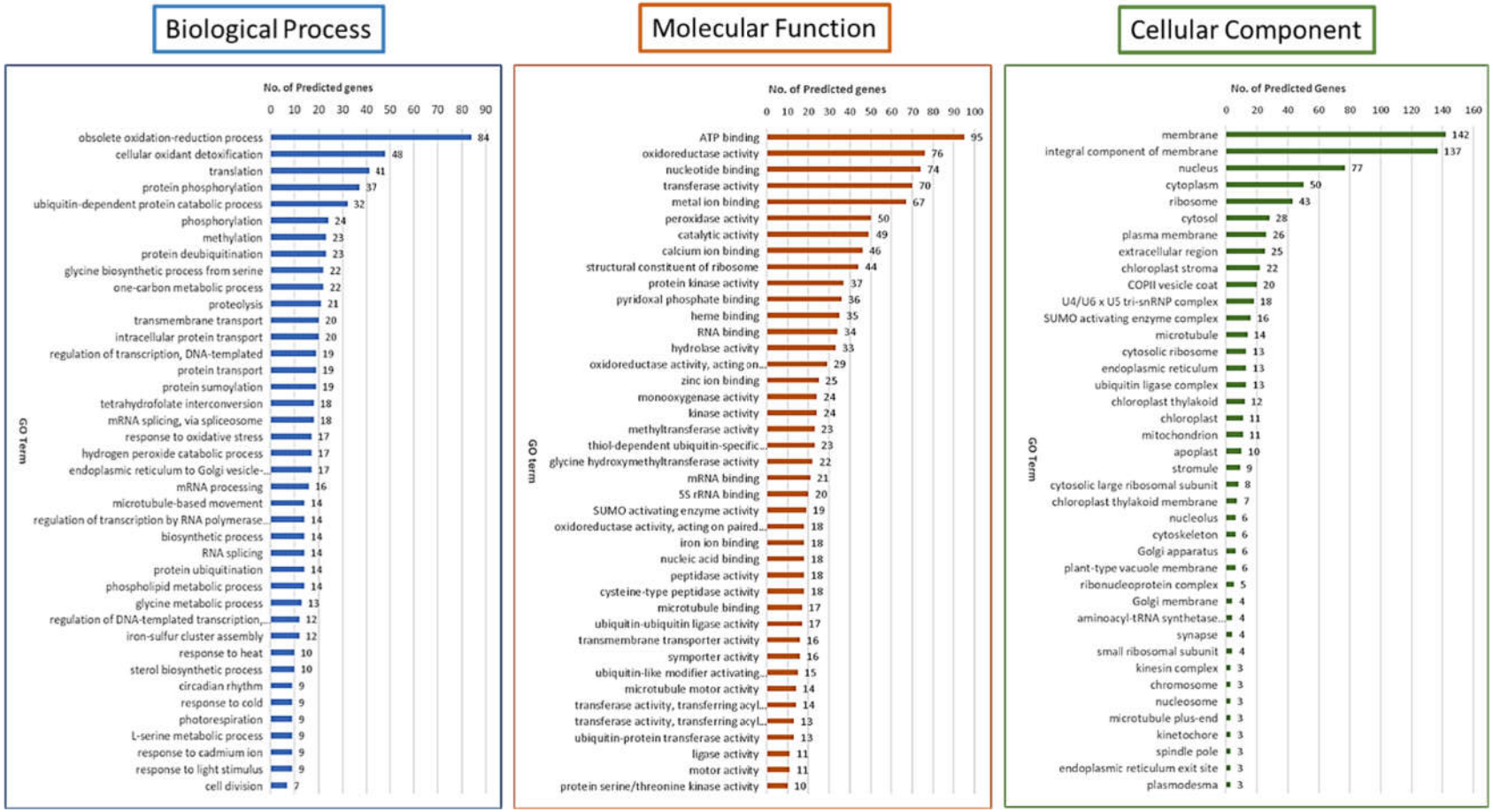


Figure 17 : GO annotations distribution for the 128 SuperTranscripts assigned by blast2go. Plot with blue bars is for biological processes, orange for molecular functions and green for cellular components. The top 40 GO distributions are shown in this figure.

5.6 BUSCO analysis & Alignment check

To test the SuperTranscriptome assembly for completeness, we used the GenomeQC[97] web application's Benchmarking Universal Single-Copy Orthologs (BUSCO)[98] implementation to search for conserved orthologous genes. GenomeQC was set to use the BUSCO dataset embryophyte_odb9(plants) along with AUGUSTUS[99] species "maize", and the option to compare the results against the precomputed results of three maize reference genomes – "MaizeB73_v4_scaffolds", "MaizeMo17_CAU_scaffolds", and "MaizeW22_NRgenes_con". We also compared the BUSCO results for the initial trinity transcript assembly, and the results of transcript redundancy reduction by CD-HIT-EST. As the goal was to determine if any unique genes or transcripts were present in the FR697 genotype compared to B73, we replaced the maize ID annotated SuperTranscripts with their respective complete maize genes. This dataset was analyzed by genomeQC, and the generated BUSCO results compared to the previous datasets, as presented in Figure 6. This dataset contained 42612 full maize genes in place of their corresponding annotated SuperTranscripts from the assembly along with the 5303 assembled transcripts, resulting in the "FR697 combined SuperTranscriptome". This dataset will be a valuable resource to understand the unique ability of maize roots to keep growing under drought stress, and to gain insights into the mechanisms of such adaptation in other plant species.

To check if the original goal of gaining a similar alignment rate of 90% for B73 RNA-Seq samples to the B73 reference genome had been met, a sample set of RNA-Seq reads from the FR697 dataset was aligned to the "merged" supertranscriptome. The results are shown in table 11revealing that the alignment improved and matched that for B73, thus indicating a successful assembly and that this dataset can be used for further downstream.

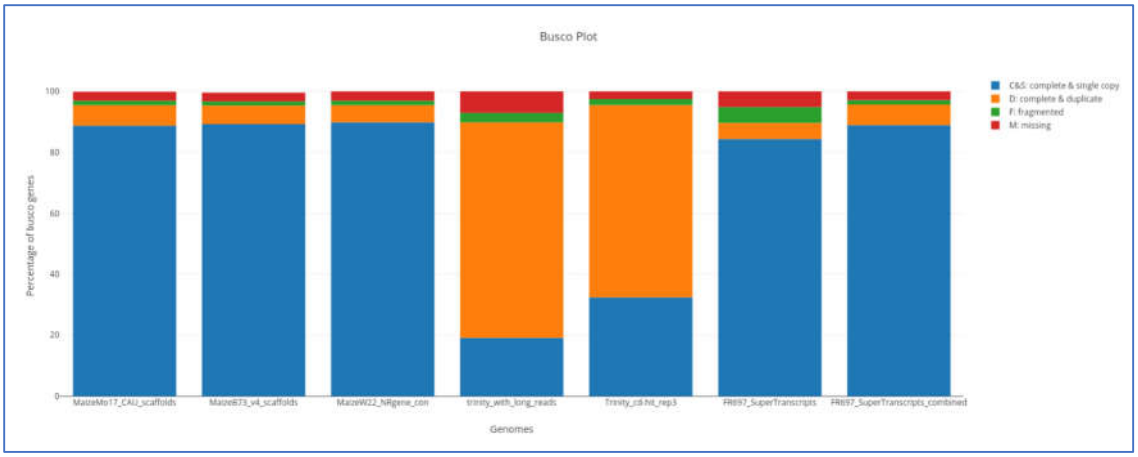


Figure 18: BUSCO analysis of merged SuperTranscriptome (full maize gene sequences replacing their annotated SuperTranscriptome counterparts) compared against the unmerged SuperTranscriptome, Trinity assembly and 3 maize reference genome datasets.

Table 11: Comparing alignments rates of merged supertranscriptome (here labelled as "FR697 Final Assembly") vs. B73 (labelled with it's genome version "AGPv4.47"). Also included are the alignment rates for the test trinity assembly only ("FR697 Test Assembly") and the original alignment rates of all the samples to an older version of B73 genome ("AGPv4.38"), used for the transcriptome analysis in chapter 2.

Sample	FR697 Test Assembly	AGPv4.38	FR697 Final assembly	AGPv4.47	Treatment
S3011	77.88%	88.45%	82.81%	**89.12%**	SSa_B73
S2022	79.80%	92.00%	85.39%	**92.69%**	SSb_B73
S4013	77.19%	88.56%	82.76%	**90.03%**	SSc_B73
S5041	80.04%	90.60%	85.05%	**91.84%**	Wwa_B73
S1012	80.51%	91.32%	86.02%	**92.98%**	WWb_B73
S6013	78.55%	88.36%	83.72%	**91.38%**	WWc_B73
S1021	**87.39%**	81.52%	**88.61%**	82.73%	SSa_FR697
S2042	**90.56%**	79.48%	**91.64%**	80.71%	SSb_FR697
S5013	**87.41%**	78.50%	**88.66%**	81.05%	SSc_FR697
S3022	**87.79%**	78.98%	**88.86%**	81.08%	WWb_FR697
S4021	**90.31%**	83.41%	**91.67%**	84.29%	WWa_FR697
S2013	**82.01%**	88.62%	**86.50%**	89.24%	WWc_FR697

5.7 Conclusion

This chapter detailed the assembly of a FR697 transcriptome using the concept of assembling and merging transcripts into sequences comparable to gene structures, called supertranscripts. The pipeline involved used a combination of FR697 RNA-Seq datasets from nodal root samples along with a dataset of iso-seq long reads. The resultant supertranscriptome reported the presence of 5303 new assembled sequences (supertranscripts), of which 128 were shown to code for proteins from the Viridiplantae proteome, thus becoming candidates for new "genes" or unknown isoforms. A sample set of RNA-Seq samples aligned to the new assembly demonstrated that the alignment rate for FR697 samples was similar to that of B73 samples aligned to the reference B73 genome. This dataset will be used to reanalyze all FR697 samples for gene expression and multiomics studies in the future.

Chapter 6: Developing visualization methods for multiomics data

The core idea behind systems biology is to incorporate all known information from interacting biological systems. Reduction strategies are designed to help researchers make sense of the massive amounts of information resulting from the marriage of so many datasets; however, this comes with the risk of missing key unknown elements. One way to overcome this is to use visualizations connected to such database frameworks which can not only utilize stored information gleaned from experimental data, but also incorporate publicly available datasets and annotations to add value to such information.

Such visual representations help generate new hypothesis by leveraging the experience and knowledge of researchers to quickly identify elements of interest. Interactive options also allow for interesting elements to be selected and their interconnection information to be downloaded for further investigation or to even conduct completely new studies which can be incorporated back into the databases. Prototypes build with two selected visualization methods in Rshiny and along with initial scripts

6.1 Collect analyzed data in an integrated database with annotations

KBCommons[46] is a unique online platform which allows users to deposit and store omics datasets generated from various experiments. It also incorporates some various bioinformatics tools to either generate new figures and charts, or to re-analyze the datasets by conducting new experiments such as differential expression analysis, etc. Specific

visualizations can also be used to download information about interactions along with lists of interesting elements. These features provide users the ability interrogate the datasets they upload and connect to public information such as annotations from databases like STRING, STITCH and KEGG. KBCommons is also extensible, i.e., species-specific information can be added as annotations which can be used to generate either another layer

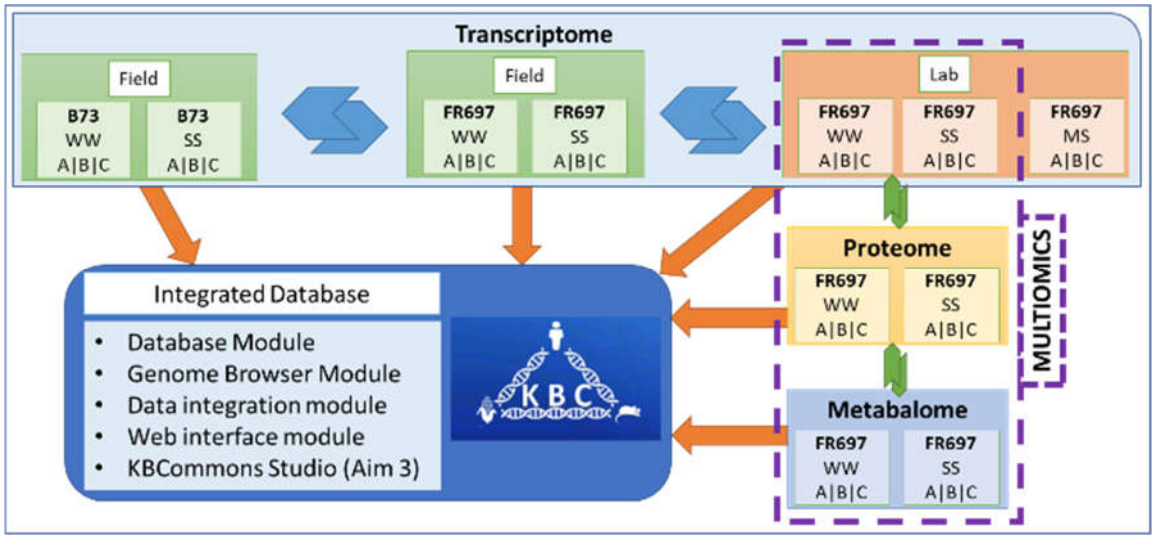


Figure 19: Multiomics datasets uploaded to a KBcommons database, such that the interactions and relationships that can be analyzed and visualized by various modules.

of visualization or highlight specific elements.

The datasets generated from the informatics studies conducted in chapter 3 and 4, have been uploaded to a specific database in the KBCommons website. The FR697 de-novo transcriptome assembly generated from the pipeline described in chapter 5 has been depThis is to allow for quick interrogation of the large number of results generated from the experiments (Figure 19). Users can query the database to filter the information based upon various parameters. This will help support their needs and questions and be used to generate new hypothesis. It also allows for the data to be reused in studies other than those

related to drought stress, expanding the scope of information that can be gleaned from these results.

6.2 Three-dimensional visualization for multiomics datasets

The Mixomics framework discussed in chapter 4 has the option to generate multiple types of visualizations. These include figures such as circle plots, Circos plots, clustered image maps (heatmaps), and even some basic 3D PCA plots which can showcase the relationships between the various elements. However, these do not allow for incorporation of external annotations without significant changes to the programming code. Even then, these changes can only be made for specific conditions or with limited scope. This is a major impediment for a broad examination of how the major elements contribute to drought stress related adaption. Thus, the results generated from multiomics pipeline were incorporated as features in the KBCommon database backend, to allow users to query and visualize based upon the components or clusters generated by the DIABLO method. Based upon these annotations, elements can be mapped to multiple pathways, which suggest that these pathways, or at least part of were active during root elongation specific processes.

To help interrogate this new type of information, a collection of RGL based visualization[43], [100] functions is being used as the base for a suite of new interactive visualization tool, called the "KBCommons Omics Studio". RGL visualizations are designed to work with processed datasets such as the ones we generated from differential expression and multiomics analysis. This also simplifies the method's extensible features, by building upon functions which are already designed to talk to similar datasets, allowing the easy addition of new visualization layers of by converting the results of any new

analysis to coordinate based datasets which can easily interact with existing ones. This also includes the option of connecting biological evidence or updated annotations provided by community databases such as MaizeKB, Phytozome, Gramene[101], etc.

The user accessible layer of visualization is implemented using the three.js[102]

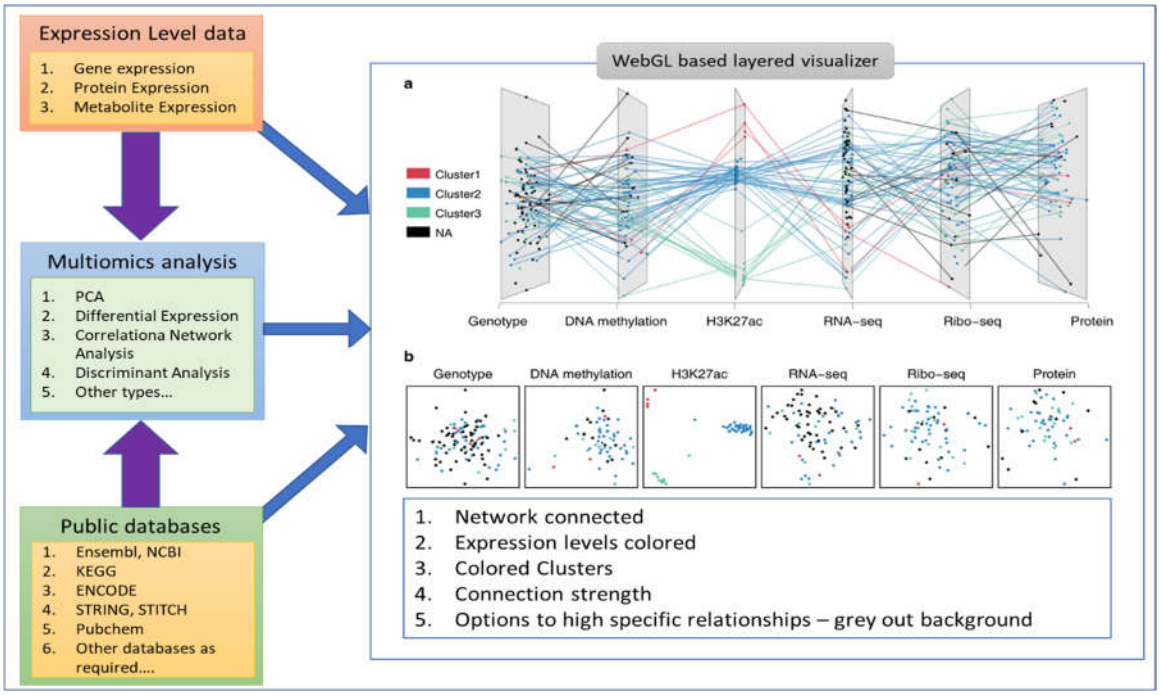


Figure 20: A schematic of how information from the multiomics analysis and annotation information contribute to the visualization method and are organized within the KBCommons database.

webGL[103] interface library to plot the 2d coordinates which extended to 3D space. The interconnections between these layers will take the information from the one-to-one mapping tables. Coordinates for the current visualization will be held in a node.js based database being used a temporary storage, which when saved by the user, will be offloaded onto the server backend of KBCommons which is deployed on the Laravel PHP web framework (Figure 20). This is technically a scale free network, so scores which define the distance between the layers will not be included. However, if certain elements from

background annotations are to be included in the visualization and were not present during the initial correlation/PCA plot generation, they will be added on to their connected element via Force directed topology[104].

The front end would allow for specific interactions to be highlighted, based upon what background information is available. For e.g., highlight only elements of DIABLO component 2 over the omics layers. This will highlight connections with respect to the position of the component over various annotations and subnetworks. Users also have the option to select specific elements and directly view their properties, it is expression values for a gene in a genotype layer. Suggested unique feature will include functions to select unconnected elements and add user annotations.

6.3 Selecting backend software and visualization packages

Various bioinformatics specific applications have been built using Rshiny. Some of these applications allow users to upload RNA-Seq datasets and perform simple gene expression analysis on them, like the START application[105]. A more advanced tool - IDEP 9.1[106] also incorporates some clustering algorithms for expanded usability. DEBrowser[107] is another such tool implemented in Rshiny and is essentially a plotting application for analyzed data. However, none of these applications are designed to handle and visualize multiple omics datasets and do not persistently store the analyzed datasets.

KBCommons Omics studio is being implemented using Rshiny, which helps build stand-alone web applications and imbed plots from R packages. A RShiny app consists of two code blocks, usually called the User Interface block (UI) and server block. As the name suggests, the server code block usually works on the server side of the webpage, and

generally most of the instructions for heavy computations are included in this section. It also controls the flow of data, including code segments which communicate with SQL databases. The UI block contains code blocks which define the characteristics of various interactive elements for e.g., slider bars, check boxes, search boxes, etc. The UI block also includes instructions on where to place these interactive elements on the screen and gives options for multi-tab webpages as well. Essentially, it is the section which defines how the webpage will look, while the server block handles all the heavy computing.

To implement the layers of interconnected plots, the GRIMONS package[43] was chosen and subsequently modified to work with Rshiny, by using R's RGL package. The efficacy

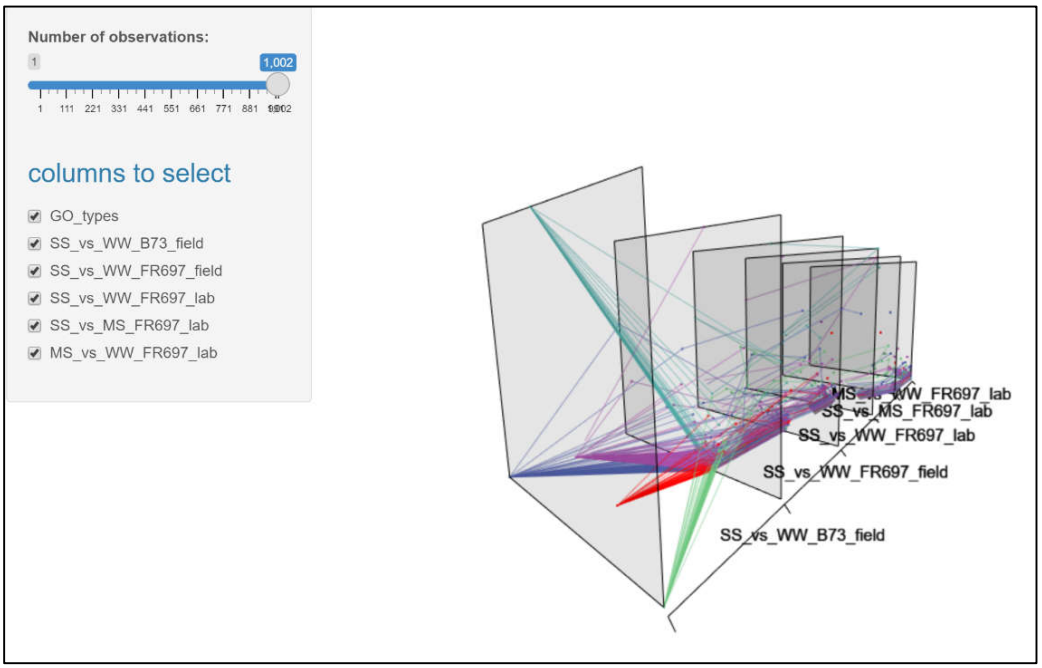


Figure 21: Basic Rshiny application with an implementation of the modified 3D plot. Layers correspond to differential expression comparisons made with Severe stressed vs Well watered plants (all tip regions merged) as listed in the options panel in the figure. A truncated list of 1000 genes were used to generate this plot. The genes were enriched with GO terms. The first layer shows GO terms, which were assigned X and Y coordinates by calculating the median of X-Y coordinates for genes associated with the term.

of this application was tested using a sample set of gene expression datasets. The

application was initialized within R as a standalone plot, and then imbedded in an Rshiny application. GRIMONS accepts a matrix of X and Y coordinates as input, where the row names define the interconnections. Columns are in pairs – denoting X and Y coordinates for each layer of the plot. The number of pairs determines how many layers are visualized. For e.g., in figure xx an example matrix – containing up to a thousand genes, and their log fold change values taken as X coordinates for each layer and Q-value (or FDR) as Y coordinates. If the element does not exist in a certain layer, NA values are assigned to the X and Y columns.

Since the intention is to make KBCommons Omics Studio a comprehensive suite of interactive visualizations for multiomics datasets, other similar interactive plot packages were also considered for inclusion. One promising package is called the Volcano3D [108], [109]. It consists of a 3-dimensional plotting device, which projects differentially expressed genes from 3 experiments, and gives users the ability to select clusters in the plot. The functions are built into the plotting device and do not need any Rshiny control widgets to interact with them.

6.4 Application prototypes and preprocessing

The implementation and functionality of GRIMONS and Volcano3D with the multiomics datasets was tested by building prototypes of Rshiny applications. A set of R scripts were also made to process the datasets into data frames compatible with the functions of GRIMONS and Volcano3D. These processed datasets are used as inputs for the prototype applications. Figure 22 showcases the prototype of a GRIMONS based implementation, including multiple filter options. Options includes remove or adding layers of omics

datasets, filtering out not significantly expressed genes from selected layer, highlighting genes which map to certain KEGG pathways and coloring according to groups of significantly expressed genes. The application seemed to get slower as more options were being selected. This was found to be a major issue through out the development of these prototypes. This is mainly because the input data frames started to become exponentially larger as more options were selected, resulting in more interconnections being processed

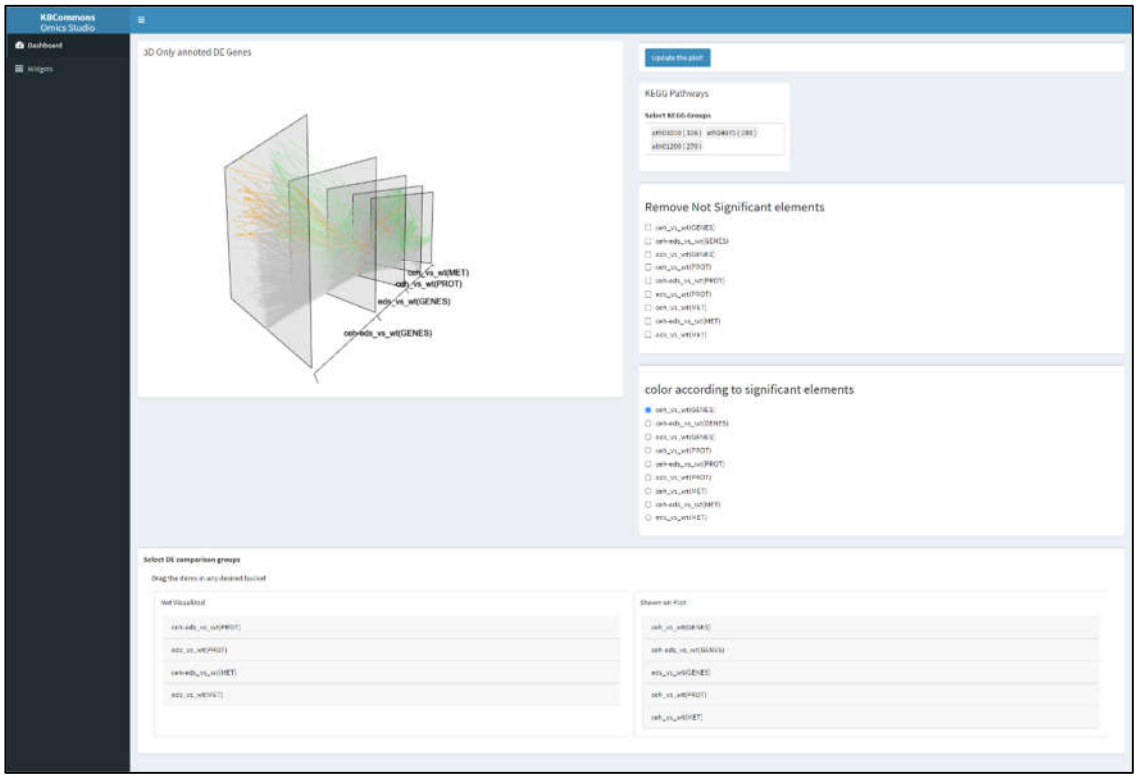


Figure 22:Screenshot of GRIMONS Rshiny prototype with multiple interactive options.

and added to these data frames. Potential solutions are to use an SQL or SQL like database and implement SQL queries directly into the processing code. This will allow the back-end code (the "server" section of Rshiny scripts) to build the data frames in real time, instead of relying of large data frame with redundant data, reducing overheard in processing time.

Figure 23 shows a prototype of Volcano3d with basic Rshiny options. It includes the option to visualize the 3 different differential expression datasets as individual volcano plots,

along with the combined 3D version. It also shows a combined box plot for the expression levels of any gene selected on the 3D visualization section. Options to select datasets, change color scheme, filter out not-significant genes and change the plot structures according to selected values such as fold change, q-value, Z-score, etc. So, no major issue was found while implementing this function, as it is a dedicated to showcase interactions between 3 DE datasets.

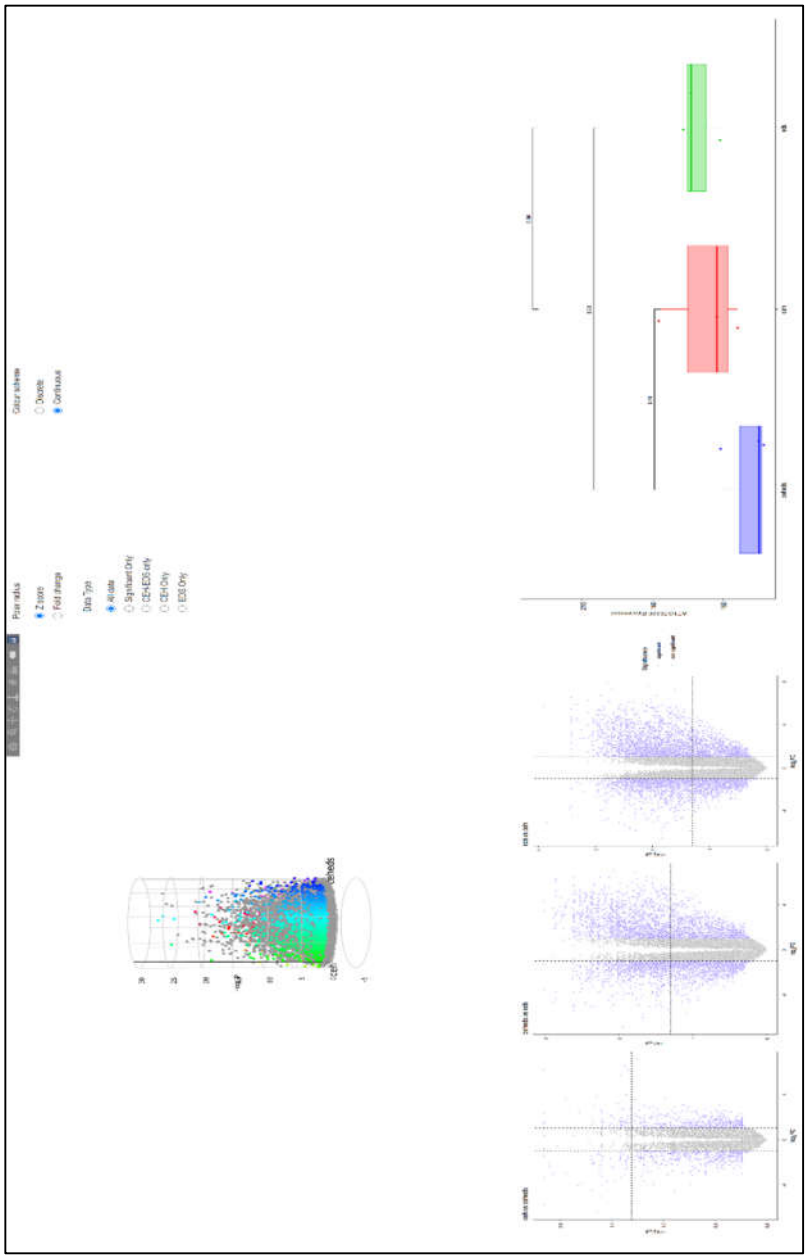


Figure 23: Screenshot of Volcano3D Rshiny prototype along with standard 2d volcano plots for each dataset along and example boxplot showing the expression levels of selected gene over the 3 datasets.

6.5 Challenges and future directions for method implementations

The development of KBCommons omics studio faces many challenges, primary of which is dealing with the large sizes of the datasets created by the addition of various annotation information. While KBCommons does have a SQL database to help manage and store the data, including access to SQL queries for quick information retrieval, there is still a considerable delay while processing the datasets into a format recognizable by the various 3D packages being tested. Some of the packages being tested also have random bugs and visualization errors due to incompatibility with the newer RGL display package. Different solutions are being evaluated, including building a brand new package dedicated to visualizing multilayered plots similar to GRIMONS, but with much more extended functionality, possibly using the rayshader[110] and rayrender[111] packages.

Eventually the Omics studio will go live and will include the multiomics data from chapters 3 and 4, presenting unique opportunities to generate more novel hypothesis based on the various new plots and images. Omics studio will also work with other datasets including raw RNA-Seq data, with the pre-processing being carried by a gene expression analysis PGen pipeline, which is part of the KBCommons suite of tools.

Chapter 7: Summary and way forward

7.1 Updating gene expression results

The RNA-Seq data used in Chapter 3 was generated from maize plant samples grown in the year 2017 as part of the drought root project by the Sharp lab. This project included plans to generate 2 more batches of data over the next two years, and both have been collected and processed. This gives the unique opportunity to not only add more replicates to the differential gene expression study, but also to compare between results between the 3 batches. However, to do this the informatics pipeline will have to be updated to use a software package which accounts for the variability in expression levels which inadvertently is always introduced when using samples generated in different batches. The effect is known as "batch effect" and various methods have been designed to reduce variation across samples and tease out the most significant differentially expressed genes from these studies.

EdgeR is a differential gene expression analysis tool which estimates dispersion of counts of aligned RNA-Seq reads per gene over replicates and normalizes them over the entire dataset of genes[112]. EdgeR also implements a generalized linear model[113] tuned by a design matrix, to normalize across multiple groups between batches, usually to account for different observations or phenotypes, such as in the case of the nodal root samples – replicates from either A, B or C root tip region, or if from B73 or FR697 genotype samples. This makes edgeR a suitable replacement for cuffdiff which unfortunately is not suited for dealing with batch effects in samples. EdgeR accepts a matrix of absolute counts of RNA-

Seq reads aligned to gene models. For this the samples will have to be realigned using a counts specific tool such as HT-seq[114] which reports how many reads align to a genomic feature.

This also opens up the opportunity to update the alignments, i.e., use a newer version of the reference B73 genome along with the FR697 de-novo transcriptome assembly, detailed in chapter 5. This will not only enable an accurate comparison of gene expression patterns across genotypes but will also update the multiomics analysis with a more robust set of significant genes. The WGCNA clustering and TF-gene relationships will also be recalculated using the new datasets.

7.2 Incorporate protein quantification datasets into the multiomics study

As mentioned before the FR697 nodal root samples were collected with the intent of generating proteomics samples as well. These samples are in the process of being quantified and analyzed to reveal proteins which are differentially accumulated in these samples. The multiomics analysis pipeline was developed with the intention of using this dataset as well, i.e, a total of four FR697 specific datasets – phenotypic observations, gene expression values, protein quantification levels and metabolite quantities. The design matrix will be updated to reflect the relationship between the 4 datasets and the method will have to be retuned, along with selecting a new distance penalty.

7.3 Merge results from both exploratory strategies

To gain deeper insights into the mechanisms behind the adaptation for drought stress, the significant elements from the multiomics study will be marked on the network visualized in Cytoscape with the TF-gene interactions. This merged network can showcase systemic

links and when connected to biochemical pathways generate unique candidates. The connected dataset will also be uploaded to KBCommons studio with options to filter according to various pathways and cutoff values according to expression levels. Other method's such as the IMPres-Pro dynamic programming-based tool[27] will also be explored to determine optimum links between various TF-Gene pairs to add more support to the results of the exploratory analysis.

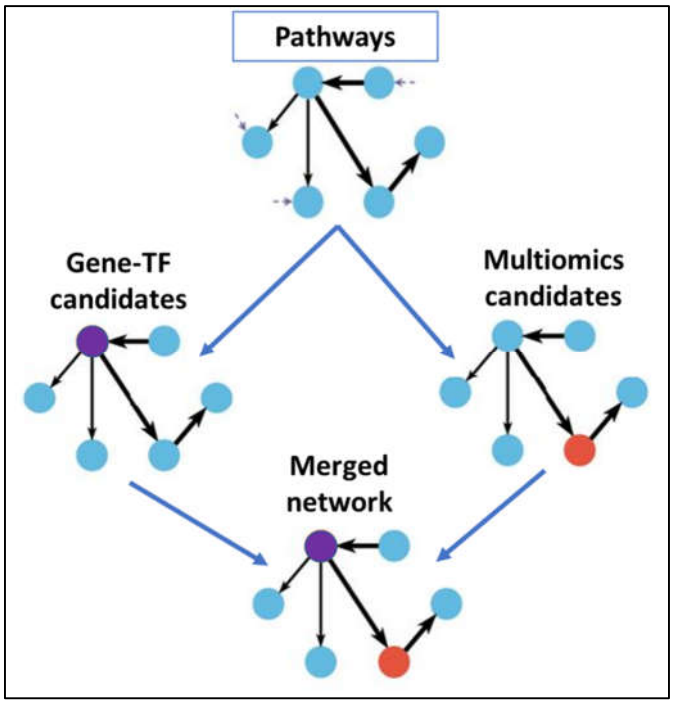


Figure 24: Example of connecting significant elements from both exploratory pipelines.

7.4 Conclusion

The exploratory strategies discussed in this dissertation were developed to explore maize samples, however they can also be used in conjunction with datasets from other projects as well. This is dependent upon processing the samples into the accepted input formats for each type of analysis. The significant elements detected at the end of these strategies will be validated and explored with laboratory methods such as qRT-PCR and hopefully reveal insights into the unique drought adaptation of maize roots.

Bibliography

[1] P. Ranum, J. P. Peña-Rosas, and M. N. Garcia-Casal, "Global maize production, utilization, and consumption," *Ann. N. Y. Acad. Sci.*, vol. 1312, no. 1, pp. 105–112, 2014.

[2] "U.S. Crops and Livestock in Drought | Drought.gov." [Online]. Available: https://www.drought.gov/drought/data-gallery/us-crops-and-livestock-drought. [Accessed: 18-Jun-2019].

[3] M. A. Ahmed, M. Zarebanadkouki, F. Meunier, M. Javaux, A. Kaestner, and A. Carminati, "Root type matters: measurement of water uptake by seminal, crown, and lateral roots in maize," *J. Exp. Bot.*, vol. 69, no. 5, pp. 1199–1206, Feb. 2018.

[4] R. E. Sharp and W. J. Davies, "Solute regulation and growth by roots and shoots of water-stressed maize plants," *Planta*, vol. 147, no. 1, pp. 43–49, Oct. 1979.

[5] M. E. Westgate and J. S. Boyer, "Osmotic adjustment and the inhibition of leaf, root, stem and silk growth at low water potentials in maize," *Planta*, vol. 164, no. 4, pp. 540–549, Jul. 1985.

[6] R. T. Clark *et al.*, "High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development," *Plant, Cell Environ.*, 2013.

[7] R. E. Sharp, W. K. Silk, and T. C. Hsiao, "Growth of the maize primary root at low water potentials : I. Spatial distribution of expansive growth.," *Plant Physiol.*, vol. 87, no. 1, pp. 50–7, May 1988.

[8] P. Voothuluru *et al.*, "Apoplastic Hydrogen Peroxide in the Growth Zone of the Maize Primary Root. Increased Levels Differentially Modulate Root Elongation Under Well-Watered and Water-Stressed Conditions," *Front. Plant Sci.*, vol. 11, p. 392, Apr. 2020.

[9] K. J. Riggs, "Maize nodal root growth under water deficits," University of Missouri, 2016.

[10] K. A. Leach, L. G. Hejlek, L. B. Hearne, H. T. Nguyen, R. E. Sharp, and G. L. Davis, "Primary Root Elongation Rate and Abscisic Acid Levels of Maize in Response to Water Stress," *Crop Sci.*, vol. 51, no. 1, pp. 157–172, Jan. 2011.

[11] T. G. Dowd, D. M. Braun, and R. E. Sharp, "Maize lateral root developmental plasticity induced by mild water stress. I: Genotypic variation across a high-resolution series of water potentials," *Plant. Cell Environ.*, vol. 42, no. 7, pp. 2259–2273, Jul. 2019.

[12] A. Asaro, B. P. Dilkes, and I. Baxter, "Multivariate analysis reveals environmental and genetic determinants of element covariation in the maize grain ionome," *bioRxiv*, p. 241380, Dec. 2017.

[13] Q. Orozco-Ramírez, J. Ross-Ibarra, A. Santacruz-Varela, and S. Brush, "Maize

diversity associated with social origin and environmental variation in Southern Mexico," *Heredity (Edinb).*, vol. 116, no. 5, pp. 477–484, May 2016.

[14] T. W. Reynolds, S. R. Waddington, C. L. Anderson, A. Chew, Z. True, and A. Cullen, "Environmental impacts and constraints associated with the production of major food crops in Sub-Saharan Africa and South Asia," *Food Secur.*, vol. 7, no. 4, pp. 795–822, Aug. 2015.

[15] J. M. McGrath *et al.*, "An analysis of ozone damage to historical maize and soybean yields in the United States.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 46, pp. 14390–5, Nov. 2015.

[16] A. S. Davis, J. D. Hill, C. A. Chase, A. M. Johanns, and M. Liebman, "Increasing Cropping System Diversity Balances Productivity, Profitability and Environmental Health," *PLoS One*, vol. 7, no. 10, p. e47149, Oct. 2012.

[17] "Roots in Drought | Split-Root System." [Online]. Available: https://rootsindrought.missouri.edu/?p=145. [Accessed: 02-Sep-2020].

[18] P. Langfelder *et al.*, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[19] D. Szklarczyk *et al.*, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.

[20] B. Snel, "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene," *Nucleic Acids Res.*, vol. 28, no. 18, pp. 3442–3444, Sep. 2000.

[21] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[22] A. Fabregat *et al.*, "Reactome pathway analysis: A high-performance in-memory approach," *BMC Bioinformatics*, vol. 18, no. 1, Mar. 2017.

[23] S. Huang, K. Chaudhary, and L. X. Garmire, "More Is Better: Recent Progress in Multi-Omics Data Integration Methods," *Front. Genet.*, vol. 8, p. 84, Jun. 2017.

[24] R. Haas, A. Zelezniak, J. Iacovacci, S. Kamrad, S. Townsend, and M. Ralser, "Designing and interpreting 'multi-omic' experiments that may change our understanding of biology," *Curr. Opin. Syst. Biol.*, vol. 6, pp. 37–45, Dec. 2017.

[25] N. Rappoport and R. Shamir, "Multi-omic and multi-view clustering algorithms: review and cancer benchmark," *Nucleic Acids Res.*, vol. 46, no. 20, pp. 10546–10562, Nov. 2018.

[26] H. R. Frost and C. I. Amos, "A multi-omics approach for identifying important pathways and genes in human cancer," *BMC Bioinformatics*, vol. 19, no. 1, p. 479, Dec. 2018.

[27] Y. Jiang, Y. Liang, D. Wang, D. Xu, and T. Joshi, "IMPRes: Integrative

MultiOmics pathway resolution algorithm and tool," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 2260–2260.

[28] H. Qin, T. Niu, and J. Zhao, "Identifying Multi-Omics Causers and Causal Pathways for Complex Traits," *Front. Genet.*, vol. 10, p. 110, Feb. 2019.

[29] P. Rinaudo, S. Boudah, C. Junot, and E. A. Thévenot, "biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data.," *Front. Mol. Biosci.*, vol. 3, p. 26, 2016.

[30] I. S. L. Zeng and T. Lumley, "Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science)," *Bioinform. Biol. Insights*, vol. 12, p. 117793221875929, Jan. 2018.

[31] Y. El-Manzalawy, "CCA based multi-view feature selection for multi-omics data integration," in *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2018, pp. 1–8.

[32] M. G. Grabherr *et al.*, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, Jul. 2011.

[33] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.

[34] M. Ramos *et al.*, "Software for the Integration of Multiomics Experiments in Bioconductor.," *Cancer Res.*, vol. 77, no. 21, pp. e39–e42, 2017.

[35] J. Ou and L. J. Zhu, "trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data," *Nat. Methods*, vol. 16, no. 6, pp. 453–454, Jun. 2019.

[36] S. D. McCabe, D.-Y. Lin, and M. I. Love, "MOVIE: Multi-Omics Visualization of Estimated contributions," *bioRxiv*, p. 379115, Jul. 2018.

[37] H. Fanaee-T and M. Thoresen, "Multi-insight visualization of multi-omics data via ensemble dimension reduction and tensor factorization," *Bioinformatics*, vol. 35, no. 10, pp. 1625–1633, May 2019.

[38] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks.," *Genome Res.*, vol. 13, no. 11, pp. 2498–504, Nov. 2003.

[39] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," *Icwsm*, pp. 361–362, 2009.

[40] M. Krzywinski *et al.*, "Circos: An information aesthetic for comparative genomics," *Genome Res.*, vol. 19, no. 9, p. 1639, Sep. 2009.

[41] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "STITCH 5: augmenting protein–chemical interaction networks with tissue and

affinity data," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D380–D384, Jan. 2016.

[42] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Res.*, vol. 36, no. Database, pp. D684–D688, Dec. 2007.

[43] M. Kanai, Y. Maeda, and Y. Okada, "Grimon: graphical interface to visualize multi-omics networks," *Bioinformatics*, vol. 34, no. 22, p. 3934, Nov. 2018.

[44] T. Joshi *et al.*, "Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics," *BMC Genomics*, vol. 13, no. Suppl 1, p. S15, Jan. 2012.

[45] T. Joshi *et al.*, "Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1245–D1252, Jan. 2014.

[46] S. Zeng, Z. Lyu, S. R. K. Narisetti, D. Xu, and T. Joshi, "Knowledge Base Commons (KBCommons) v1.0: A multi OMICS' web-based data integration framework for biological discoveries," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 589–594.

[47] L. M. Dwyer, D. W. Stewart, L. Carrigan, B. L. Ma, P. Neave, and D. Balchin, "Guidelines for Comparisons among Different Maize Maturity Rating Systems," *Agron. J.*, vol. 91, no. 6, pp. 946–949, Nov. 1999.

[48] "PacificBiosciences/IsoSeq: IsoSeq3 - Scalable De Novo Isoform Discovery from Single-Molecule PacBio Reads." [Online]. Available: https://github.com/PacificBiosciences/IsoSeq. [Accessed: 03-Mar-2021].

[49] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Studying Gene Expression and Function," 2002.

[50] F. Katagiri and J. Glazebrook, "Overview of mRNA expression profiling using DNA microarrays," *Current Protocols in Molecular Biology*, vol. Chapter 22, no. SUPPL. 85. Curr Protoc Mol Biol, 2009.

[51] R. Bumgarner, "Overview of dna microarrays: Types, applications, and their future," *Curr. Protoc. Mol. Biol.*, vol. 0 22, no. SUPPL.101, p. Unit, 2013.

[52] M. Yamamoto, T. Wakatsuki, A. Hada, and A. Ryo, "Use of serial analysis of gene expression (SAGE) technology," *J. Immunol. Methods*, vol. 250, no. 1–2, pp. 45–66, Apr. 2001.

[53] M. Hu and K. Polyak, "Serial analysis of gene expression," *Nat. Protoc.*, vol. 1, no. 4, pp. 1743–1760, Nov. 2006.

[54] A. Conesa *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1. BioMed Central Ltd., pp. 1–19, 26-Jan-2016.

[55] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1. Nature Publishing Group, pp. 57–63, Jan-2009.

[56] C. Trapnell *et al.*, “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks,” *Nat. Protoc.*, vol. 7, no. 3, pp. 562–578, Mar. 2012.

[57] Simon Andrews, “Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data,” *2010*. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed: 10-Feb-2017].

[58] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data.,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–20, Aug. 2014.

[59] Y. Jiao *et al.*, “Improved maize reference genome with single-molecule technologies,” *Nature*, vol. 546, no. 7659, pp. 524–527, Jun. 2017.

[60] D. Kim, B. Langmead, and S. L. Salzberg, “HISAT: a fast spliced aligner with low memory requirements.,” *Nat. Methods*, vol. 12, no. 4, pp. 357–60, Apr. 2015.

[61] C. Trapnell *et al.*, “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.,” *Nat. Protoc.*, vol. 7, no. 3, pp. 562–78, Mar. 2012.

[62] Z. Du, X. Zhou, Y. Ling, Z. Zhang, and Z. Su, “agriGO: a GO analysis toolkit for the agricultural community.,” *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W64-70, Jul. 2010.

[63] M. Kanehisa *et al.*, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[64] J. Mariette and N. Villa-Vialaneix, “Unsupervised multiple kernel learning for heterogeneous data integration,” *Bioinformatics*, vol. 34, no. 6, pp. 1009–1015, Mar. 2018.

[65] P. Shannon *et al.*, “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks,” *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[66] S. Huang, K. Chaudhary, and L. X. Garmire, “More Is Better: Recent Progress in Multi-Omics Data Integration Methods,” *Front. Genet.*, vol. 8, no. JUN, p. 84, Jun. 2017.

[67] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, “Multi-omics Data Integration, Interpretation, and Its Application,” *Bioinformatics and Biology Insights*, vol. 14. SAGE Publications Inc., 2020.

[68] A. Tenenhaus, C. Philippe, V. Guillemot, K. A. Le Cao, J. Grill, and V. Frouin, “Variable selection for generalized canonical correlation analysis,” *Biostatistics*, vol. 15, no. 3, pp. 569–583, 2014.

[69] L. C. Lee, C. Y. Liong, and A. A. Jemain, “Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps,” *Analyst*, vol. 143, no. 15.

Royal Society of Chemistry, pp. 3526–3539, 07-Aug-2018.

[70] D. M. Rotroff and A. A. Motsinger-Reif, “Embracing Integrative Multiomics Approaches.,” *Int. J. Genomics*, vol. 2016, p. 1715985, 2016.

[71] L. P. Matic *et al.*, “Novel Multiomics Profiling of Human Carotid Atherosclerotic Plaques and Plasma Reveals Biliverdin Reductase B as a Marker of Intraplaque Hemorrhage,” *JACC Basic to Transl. Sci.*, vol. 3, no. 4, pp. 464–480, Aug. 2018.

[72] Z. Costello and H. G. Martin, “A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data,” *npj Syst. Biol. Appl.*, vol. 4, no. 1, p. 19, Dec. 2018.

[73] M. Santolini *et al.*, “A personalized, multiomics approach identifies genes involved in cardiac hypertrophy and heart failure,” *npj Syst. Biol. Appl.*, vol. 4, no. 1, p. 12, Dec. 2018.

[74] D. B. Gutierrez *et al.*, “An Integrated, High-Throughput Strategy for Multiomic Systems Level Analysis,” *J. Proteome Res.*, p. acs.jproteome.8b00302, Aug. 2018.

[75] S. Biyiklioglu *et al.*, “A large-scale multiomics analysis of wheat stem solidness and the wheat stem sawfly feeding response, and syntenic associations in barley, Brachypodium, and rice,” *Funct. Integr. Genomics*, vol. 18, no. 3, pp. 241–259, May 2018.

[76] F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao, “mixOmics: An R package for ‘omics feature selection and multiple data integration,” *PLOS Comput. Biol.*, vol. 13, no. 11, p. e1005752, Nov. 2017.

[77] M. Mingon Kang, B. Baoju Zhang, X. Xiaoyong Wu, C. Chunyu Liu, and J. Gao, “Sparse generalized canonical correlation analysis for biological model integration: A genetic study of psychiatric disorders,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, vol. 2013, pp. 1490–1493.

[78] E. Robotti, M. Manfredi, and E. Marengo, “Biomarkers Discovery through Multivariate Statistical Methods: A Review of Recently Developed Methods and Applications in Proteomics,” *J Proteomics Bioinform*, no. S3, 2014.

[79] J. L. Portwood *et al.*, “MaizeGDB 2018: the maize multi-genome genetics and genomics database,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1146–D1154, Jan. 2019.

[80] N. M. Davidson, A. D. K. Hawkins, and A. Oshlack, “SuperTranscripts: A data driven reference for analysis and visualisation of transcriptomes,” *Genome Biol.*, vol. 18, no. 1, p. 148, Aug. 2017.

[81] M. Guttman *et al.*, “Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs,” *Nat. Biotechnol.*, vol. 28, no. 5, pp. 503–510, May 2010.

[82] W. R. Jeck *et al.*, “Extending assembly of short DNA sequences to handle error,”

Bioinformatics, vol. 23, no. 21, pp. 2942–2944, Nov. 2007.

[83] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, "Assembling millions of short DNA sequences using SSAKE," *Bioinformatics*, vol. 23, no. 4, pp. 500–501, Feb. 2007.

[84] A. J. Nederbragt, "On the middle ground between open source and commercial software - the case of the Newbler program," *Genome Biology*, vol. 15, no. 4. BioMed Central Ltd., p. 113, 29-Apr-2014.

[85] G. Gonnella and S. Kurtz, "Readjoiner: a fast and memory efficient string graph-based sequence assembler," *BMC Bioinformatics*, vol. 13, no. 1, p. 82, May 2012.

[86] J. T. Simpson and R. Durbin, "Efficient de novo assembly of large genomes using compressed data structures," *Genome Res.*, vol. 22, no. 3, pp. 549–556, Mar. 2012.

[87] Y. Xie *et al.*, "SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads," *Bioinformatics*, vol. 30, no. 12, pp. 1660–1666, Jun. 2014.

[88] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Res.*, vol. 19, no. 6, pp. 1117–1123, Jun. 2009.

[89] S. Gnerre *et al.*, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 4, pp. 1513–1518, Jan. 2011.

[90] N. M. Davidson and A. Oshlack, "Necklace: combining reference and assembled transcriptomes for more comprehensive RNA-Seq analysis," *Gigascience*, vol. 7, no. 5, pp. 1–6, May 2018.

[91] S. McGinnis and T. L. Madden, "BLAST: At the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Res.*, vol. 32, no. WEB SERVER ISS., p. W20, Jul. 2004.

[92] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[93] A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, Sep. 2005.

[94] Y. Onda, T. Matsumura, Y. Kimata-Ariga, H. Sakakibara, T. Sugiyama, and T. Hase, "Differential interaction of maize root ferredoxin:NADP+ oxidoreductase with photosynthetic and non-photosynthetic ferredoxin isoproteins," *Plant Physiol.*, vol. 123, no. 3, pp. 1037–1045, Jul. 2000.

[95] J. Gao *et al.*, "MaMAPK3-MaICE1-MaPOD P7 pathway, a positive regulator of cold tolerance in banana," *BMC Plant Biol.*, vol. 21, no. 1, p. 97, Dec. 2021.

[96] G. Ramakrishna *et al.*, "Comparative transcriptome analyses revealed different heat stress responses in pigeonpea (Cajanus cajan) and its crop wild relatives,"

Plant Cell Rep., vol. 40, no. 5, pp. 881–898, Apr. 2021.

[97] N. Manchanda *et al.*, "GenomeQC: A quality assessment tool for genome assemblies and gene structure annotations," *BMC Genomics*, vol. 21, no. 1, Mar. 2020.

[98] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, Oct. 2015.

[99] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "AUGUSTUS: A b initio prediction of alternative transcripts," *Nucleic Acids Res.*, vol. 34, no. WEB. SERV. ISS., pp. W435–W439, Jul. 2006.

[100] "CRAN - Package rgl: 3D Visualization Using OpenGL." [Online]. Available: https://cran.r-project.org/web/packages/rgl/index.html. [Accessed: 18-Jun-2019].

[101] D. H. Ware *et al.*, "Gramene, a tool for grass genomics.," *Plant Physiol.*, vol. 130, no. 4, pp. 1606–13, Dec. 2002.

[102] "three.js – Javascript 3D library." [Online]. Available: https://threejs.org/. [Accessed: 18-Jun-2019].

[103] "WebGL: 2D and 3D graphics for the web - Web APIs | MDN." [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/WebGL_API. [Accessed: 18-Jun-2019].

[104] W. Didimo, G. Liotta, and S. A. Romeo, "Topology-Driven Force-Directed Algorithms," Springer, Berlin, Heidelberg, 2011, pp. 165–176.

[105] J. W. Nelson, J. Sklenar, A. P. Barnes, and J. Minnier, "The START App: a web-based RNAseq analysis and visualization resource," *Bioinformatics*, vol. 33, no. 3, p. btw624, Sep. 2016.

[106] S. X. Ge, E. W. Son, and R. Yao, "iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data," *BMC Bioinformatics*, vol. 19, no. 1, Dec. 2018.

[107] A. Kucukural, O. Yukselen, D. M. Ozata, M. J. Moore, and M. Garber, "DEBrowser: Interactive differential expression analysis and visualization tool for count data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems," *BMC Genomics*, vol. 20, no. 1, p. 6, Jan. 2019.

[108] "volcano3D." [Online]. Available: https://cran.r-project.org/web/packages/volcano3D/vignettes/Vignette.html. [Accessed: 23-Apr-2021].

[109] M. J. Lewis *et al.*, "Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes," *Cell Rep.*, vol. 28, no. 9, pp. 2455-2470.e5, Aug. 2019.

[110] "CRAN - Package rayshader." [Online]. Available: https://cloud.r-project.org/web/packages/rayshader/index.html. [Accessed: 23-Apr-2021].

[111] T. Morgan-Wall, "Build and Raytrace 3D Scenes [R package rayrender version 0.21.2]," Apr. 2021.

[112] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.

[113] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.," *Nucleic Acids Res.*, vol. 40, no. 10, pp. 4288–97, May 2012.

[114] S. Anders, P. T. Pyl, and W. Huber, "HTSeq--a Python framework to work with high-throughput sequencing data.," *Bioinformatics*, vol. 31, no. 2, pp. 166–9, Jan. 2015.

Vita

Sidharth Sen is from New Delhi, India. He received a Bachelor of Technology in Bioinformatics from Dr. D. Y. Patil University, Pune, India in 2011. He then went onto to earn a Master of Research in Computational Biology from University of York, UK in 2013. He joined the PhD program in the University of Missouri Institute of Data Science and Informatics in August 2014.

He worked with his advisor Dr. Trupti Joshi on his doctoral research on developing multiomics informatics pipelines to find drought related signatures in maize nodal roots. His research interests include gene expression analysis and systems biology research with a focus on multiomics data integration method development.