Differential analysis of gene expression in *Arabidopsis thaliana* organ boundary mutants

By:
Jack Mason

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of the requirements of the Sally McDonnell Barksdale Honors College.

Oxford
May 2019

Approved by:

_____
Advisor: Dr. Sarah Liljegren

_____
Reader: Dr. Brian Doctor

_____
Reader: Dr. Bradley Jones

ACKNOWLEDGEMENTS

Firstly, I would like to thank my thesis advisor, teacher, and mentor Dr. Sarah Liljegren. Your love for the sciences and passion for genetics has opened my mind to the world of research. Throughout the years in your lab, you have given me lessons that will hold their value throughout the rest of my academic career. Thank you for allowing me to be a part of your team.

To my parents, H.F. and Laura Mason, who have shown me unconditional love and support throughout my entire life. No matter what I set as my goals, you continue to stick with me and always remind me to do what I am passionate about. You are the ones that have provided me with this opportunity to learn and I could not be more thankful.

Moreover, this project was all possible thanks to the help of Kate Childers, Katherine Anderson, and all of the members of the Liljegren Lab team. It is your intelligence and drive that pushed me to do my best every day. Thank you for all of your hard work and time needed to see this project to completion.

Finally, I would like to acknowledge the Sally McDonnell Barksdale Honors College for giving me the chance to express myself in the field of academia without bounds. Entering into the program as a Junior, I never expected the acceptance and support that I have received in the past two years. Fellow members and staff of the SMBHC have provided endless resources that have allowed me to grow in more way than I can express.

This research was supported by an NSF grant (IOS-1453733) to SL, an NSF grant (DUE-1323522) to Cold Spring Harbor Laboratories, and NSF support (DBI-0735191, DBI-1265383, DBI-1743442) of CyVerse.

ABSTRACT

JACK GORDON MASON: Differential gene expression in *Arabidopsis thaliana* mutants
(Under the direction of Dr. Sarah Liljegren)

Differences in gene expression occur due to cell type differentiation, the stages of an organism's life cycle, genetic variation, and changes in the environment. The quantification of these changes can be valuable for helping deduce the molecular mechanisms underlying the development of model organisms such as *Arabidopsis*. A pair of homeobox genes, *ARABIDOPSIS THALIANA HOMEOBOX GENE1* (*ATH1*) and *SHOOT MERISTEMLESS* (*STM*), encode transcription factors that play key roles in establishing organ boundaries in *Arabidopsis* flowers. Plants that carry mutations in both of these genes are missing floral organ-stem boundary regions and fail to shed their outer floral organs. Previous studies have shown that expression of *HAESA*, a receptor-like kinase that activates a signaling cascade necessary for organ shedding, is substantially reduced in *stm ath1* mutant flowers.

RNA-Sequencing (RNA-Seq) was used to profile the transcriptomes of wild-type and *stm ath1* inflorescences (clusters of flowers). The first aim of my study was to test the hypothesis that the expression of *HAE* is reduced in *stm ath1* inflorescences compared to wild-type. The second aim was to identify other candidate genes whose expression may be regulated by the STM and ATH1 transcription factors and that may play downstream roles in forming organ boundaries and regulating organ abscission. My analysis of a pilot RNA-seq study carried out through the 'RNA-Seq for the Next Generation' project at

Cold Spring Harbor Laboratory identified a set of twenty-four genes that are expressed at significantly lower levels in *stm ath1* mutants compared to wildtype. Although expression of *HAE* is also reduced in *stm ath1* plants compared to wildtype, this result does not fall within the range considered to be significant.

TABLE OF CONTENTS:

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

ATH1	ARABIDOPSIS THALIANA HOMEOBOX GENE1
AZ	abscission zone
BP	BREVIPIDICULLUS
ddH_2O	distilled and deionized water
DNA	deoxyribonucleic acid
FU	florescence units
GO	gene oncology
GUI	graphical user interface
GUS	β- glucuronidase
HAE	HAESA
JSD	Jean-Shannon divergence
MAP	Mitogen-activated protein
MAPK	MAP kinase
MKK4	MAP kinase kinase 4
MKK5	MAP kinase kinase 5
NGS	next-generation sequencing
NLS	nuclear localization signal
PCA	principal component analysis

PCR	polymerase chain reaction
RLK	receptor-like protein kinase
RNA	ribonucleic acid
RIN	RNA integrity number
RNA-Seq	RNA Sequencing
SAM	shoot apical meristem
scRNA-Seq	single-cell RNA-Sequencing
SEM	scanning electron micrograph
STM	SHOOT MERISTEMLESS
TAIR	The Arabidopsis Information Resource
TF	transcription factor
UV	ultraviolet light
W343*	Tryptophan at position 343 changed to a stop codon
WT	wild-type
Y399*	Tyrosine at position 399 changed to a stop codon

1. INTRODUCTION:

Arabidopsis thaliana is one of the most extensively used model organisms for plant genetic research. Its advantageous characteristics include a short lifespan, easily replicable growth conditions, relatively small size, and a diploid genome of five chromosomes that was sequenced almost twenty years ago (Jinn et al., 1999; The *Arabidopsis* Genome Initiative, 2000). Due to the conservation of many genes among plant species, the studies done using *Arabidopsis* are often useful in other plant species as well (Meinke et al., 1998).

A diagram of the floral organs in an *Arabidopsis* flower can be seen in **Figure 1**. Establishment of floral organ boundaries occurs at multiple steps of flower development. Boundaries in *Arabidopsis* serve to separate functional domains of cells (Yu and Huang 2016). They allow for distinction between the flower meristem and the shoot meristem from which it originates, between floral organ primordia and the underlying stem (the floral pedicel) and between adjacent floral organs (**Figures 1 and 3**; Aida and Tasaka, 2006). Cells within organ boundaries display morphologically unique characteristics such as a smaller size and lower rate of cell division (Breuil-Broyer et al., 2004). Cells within boundary regions also appear to set the stage for differentiation of certain cell fates. For example, it is known that the floral organ-pedicel boundary regions give rise to

abscission zones (AZ) that allow the flower to shed its outer floral organs (Gubert et al., 2014). The specialized cells within abscission zones, when activated, secrete sets of enzymes that modify the cell walls and digest the middle lamella between neighboring cell layers in these zones (Rubenstein and Leopold 1964).

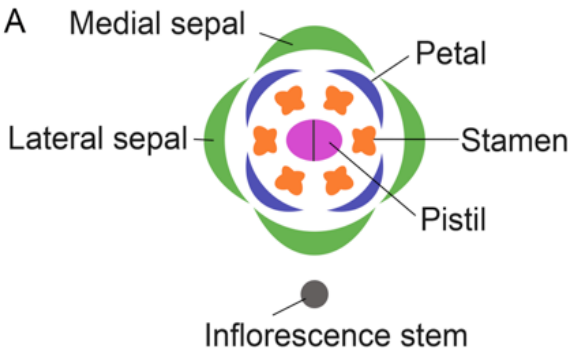

Figure 1. Illustration of the floral organs present in a wild-type *Arabidopsis thaliana* flower. Boundaries are established between adjacent floral organs.
(Image credit: Gubert et al., 2014)

As shown in **Figure 2**, two homeobox genes that contribute to the formation of floral organ boundary regions are *SHOOT MERISTEMLESS* (*STM*) and *ARABIDOPSIS THALIANA HOMEOBOX GENE1* (*ATH1*) (Long et al., 1996; Gómez-Mena et al., 2008; Raybourn, 2016; Malone, 2018; Childers, 2018). Members of this family of transcription factors have a DNA binding domain composed of sixty amino acids that is known as a homeodomain (**Figure 2**). This domain folds into a unique shape that binds to conserved sequences in target genes known as consensus sequences. Expression of the target genes can be either positively or negatively regulated by this type of transcription factor (Bürglin et al., 2015, Mukherjee et al., 2009).

ATH1 is a BELL-type homeodomain transcription factor that plays an important role in forming a boundary between the basal regions of *Arabidopsis* floral organs and the underlying floral stem, or pedicel (Gómez-Mena et al., 2008). Unlike wild-type flowers, *ath1* mutant flowers do not shed their stamens after fertilization. In addition, cross-sections of the mutant flowers showed that the small cells characteristic of the stamen abscission zones were missing and that the stamens were partially fused at their base (Gómez-Mena et al., 2008).

Scanning electron micrograph (SEM) images of *ath1* single mutant flowers show that the boundary between the sepals and underlying pedicel is less defined than in wild-type flowers (**Figure 3)**. A gene that is expressed in the pedicel and sepal-pedicel boundary region but restricted from the sepals is *BREVIPEDICELLUS (BP)* (Ori et al., 2000). A molecular marker for *BP* was created by fusing the *BP* promoter to a reporter gene that encodes the bacterial enzyme β-glucuronidase (GUS). Cells that express the *BP:GUS* transgene turn blue when they are exposed to X-Gluc, the substrate for the GUS enzyme. Compared to wild-type flowers, diminished expression of the *BP:GUS* transgene is found in the pedicels and boundary regions of *ath1* flowers (Gómez-Mena et al., 2008). Taken together, these findings suggest that ATH1 plays a role in the formation of organ boundaries.

STM encodes a KNOTTED-like homeodomain transcription factor (Long et al., 1996). As STM does not contain the nuclear localization signal (NLS) needed for entrance through the nuclear membrane, it heterodimerizes with BELL-type homeodomain transcription factors, including ATH1, to enter the nucleus and regulate its target genes (Cole et al., 2006; Rutjens et al., 2009). Studies of a hypomorphic (partial loss-of-

function) *stm* mutant (**Figure 2**) have revealed that STM also plays an important role in establishing floral organ boundaries. Like *ath1* flowers, *stm* flowers retain their stamens and their sepal-pedicel boundaries are less distinct (Palmer, 2018).

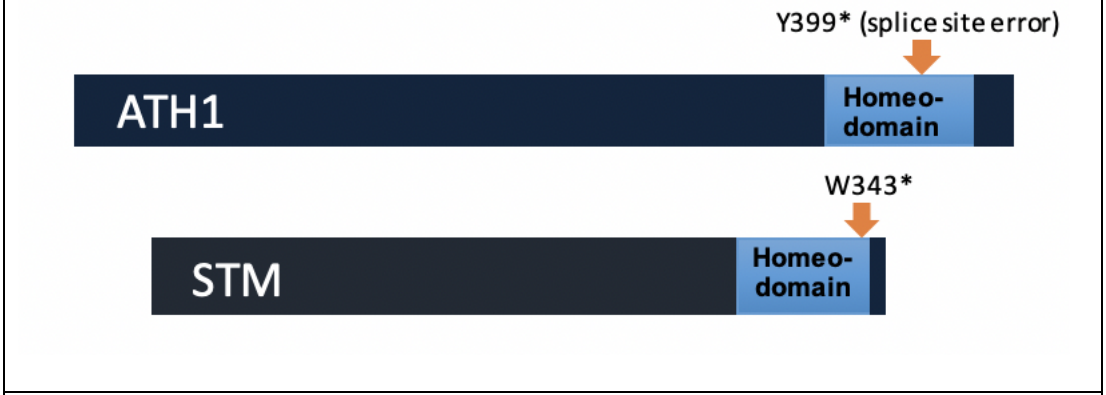

Figure 2. The ATH1 and STM transcription factors are altered by mutations within their homeodomains. The light blue boxes labeled "homeodomain" denote the DNA binding domains for this family of transcription factors**.** The *stm ath1* double mutants used in this study are homozygous for the depicted nonsense alleles.
(Image credit: Raybourn, 2016)

When plants are homozygous for both the hypomorphic allele of *STM* and a loss-of-function allele of *ATH1*, their phenotypes are enhanced relative to either single mutant. Both mutations introduce premature stop codons within the homeodomains of these transcription factors (**Figure 2)**. In addition to the stamens, abscission of the sepals and petals is blocked in *stm ath1* double mutant flowers (Palmer, 2018). Furthermore, the sepal-pedicel boundary is completely obscured in *stm ath1* flowers (**Figure 3)**. Taken together, these results suggest that by establishing floral organ boundaries, the ATH1 and STM transcription factors are setting the stage for abscission zone development.

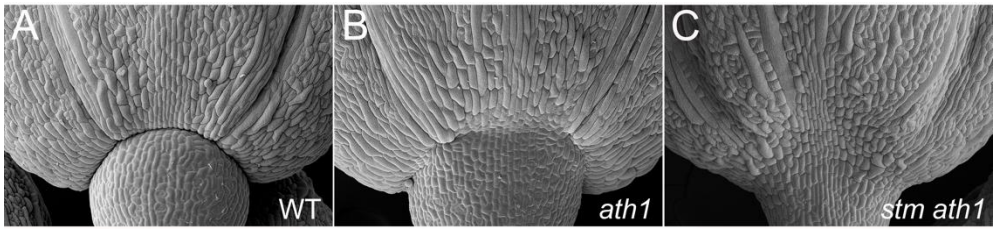

Figure 3. The sepal-stem boundary is altered in *stm ath1* flowers. Scanning electron micrographs of a wild-type and mutant *Arabidopsis* flowers.
(A) A well-defined boundary is formed between the sepals and stem of wild-type flowers.
(B) In *ath1* single mutant flowers, this organ boundary is less distinct.
(C) In *stm ath1* double mutant flowers, this boundary is absent.
(Image credit: Liljegren Lab, unpublished work)

There is much to be discovered about the transcriptional circuitry that specifies the floral organ-pedicel boundaries and how components of this circuitry may regulate the subsequent differentiation of abscission zone cells found at the bases of the sepals, petals and stamens. As shown in **Figure 4**, a candidate target of ATH1 and STM is the gene *HAESA* (*HAE*), which encodes a leucine-rich repeat receptor-like protein kinase. Kinases are enzymes that phosphorylate other proteins to either activate or deactivate them. HAE and its functionally redundant partner, HAESA-LIKE2 (HSL2) are transmembrane receptors that activate a MAPK signaling cascade required for enacting organ abscission (Jinn et al., 1999; Cho et al., 2008). Plants that carry mutations in both *HAE* and *HSL2* fail to shed their floral organs. A GUS marker controlled by a translational fusion to the *HAE* promoter has been previously used to track abscission zone cells in wild-type and mutant flowers (Leslie et al., 2010; Gubert et al., 2014). Analysis with this marker revealed a substantial reduction of *HAE:GUS* expression in the floral organ-pedicel junctions of *stm ath1* flowers compared to wildtype (**Figure 4**; Raybourn, 2016).

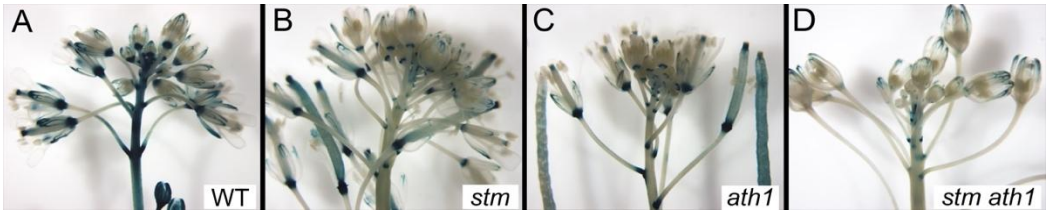

Figure 4. Expression of the *HAE:GUS* marker is significantly reduced in *stm ath1* flowers. Images of wild-type and mutant inflorescences.
(A) In wild-type plants, expression of the *HAESA* receptor-like kinase is found in the abscission zone cells at the sepal-pedicel region of each flower.
(B, C) In *stm* and *ath1*single mutant plants *HAE:GUS* expression is still observed at the sepal-pedicel boundaries.
(D) In *stm ath1* double mutant plants *HAE:GUS* expression in this region is significantly reduced.
(Image credit: Raybourn 2016)

To uncover additional candidates that may be targets of STM and ATH1 and to determine whether additional evidence supports *HAE* as a target, I have used a bioinformatic approach to analyze the transcriptional profiles of the *stm ath1* mutant and wild-type plants. My project examined the results of a pilot RNA-Sequencing (RNA-Seq) study the Liljegren lab carried out with the support of Cold Spring Harbor Laboratory's RNA-Seq for the Next Generation initiative (www.rnaseqforthenextgeneration.org). RNA-Seq is the application of next-generation sequencing (NGS) techniques to quantify the transcriptional RNA of a desired sample (Chu and Corey, 2012). As illustrated in **Figure 5**, the total RNA of a sample is converted into complementary DNA (cDNA) via reverse transcription. The cDNA library is then compared to a reference genome, and the abundance of RNA can be estimated based on calculations (Chu and Corey, 2012). This approach allows all of the expressed RNA within a cell, its transcriptome, to be quantified. Through advancements in technology and decreases in usage cost, this procedure has become a common tool in the life sciences community. RNA-Seq has

many uses ranging from solo transcriptome profiling, multi-subject comparison for genomic expression differentiation, to single-cell RNA-Seq (Conesa et al., 2016).

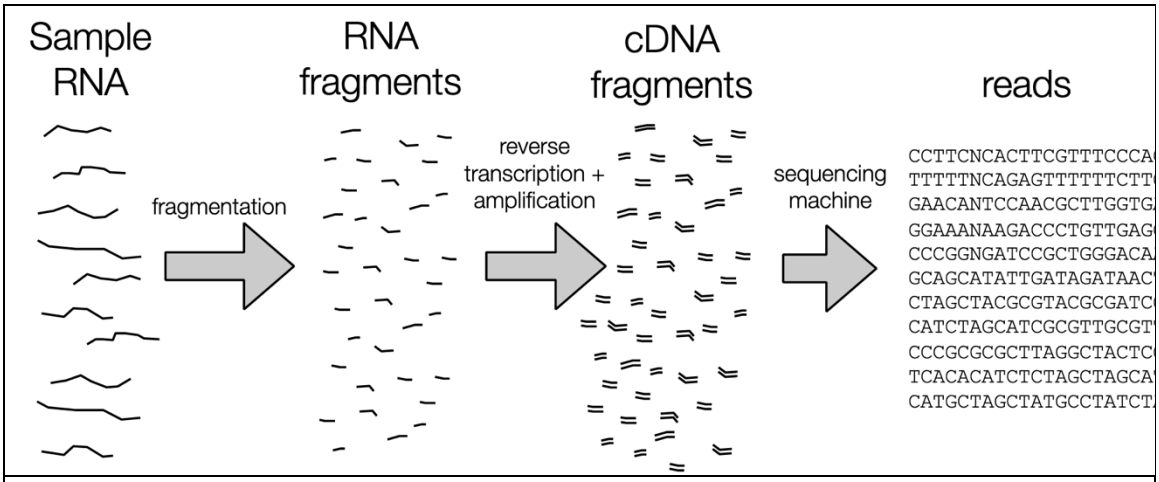

Figure 5. Overview of an RNA-Seq protocol. Sample RNA is obtained and reduced to fragments. These segments are then reverse transcribed into complementary DNA (cDNA) and sent to an off-site lab for sequencing. Raw reads are then produced and are ready for alignment to the reference genome.
(Image credit: Dewey, 2014)

The use of open-source software and public data sharing has allowed bioinformaticians to provide training to the scientific community in the use of RNA-Seq. Organizations such as CyVerse have developed user-friendly interfaces that make high-level genomic analysis widely available and easy to use (Merchant et al., 2016). The CyVerse collaborative known as DNA-Subway is a workspace that allows for DNA sequence annotations, phylogenetic analyses, and study of NGS data. Each of these workflows is illustrated as a "subway line" that directs the user through the maze of sequence analysis in a simplified fashion. Directed at increasing undergraduate education in the field of bioinformatics, DNA Subway has done just that by providing hands-on learning possibilities (Williams, 2008).

Due to the rapid evolution of technology, the most effective "pipeline" for RNA-Seq analysis is changing frequently. Pipeline is a term used in bioinformatics to

reference the series of steps taken or programs used to process data. The typical RNA-Seq pipeline involves qualification checks to ensure that the samples reads are clear enough to continue, quantification of the RNA transcriptome against a reference genome, and analysis of differential abundance. DNA Subway currently uses the quantification program known as Kallisto, which is two orders of magnitude faster than its predecessors and of similar accuracy (Bray et al., 2016). The Kallisto program takes advantage of a process known as "pseudoalignment", which speeds up the read process by not having to compare each RNA transcript read to the reference genome. Transcript reads are stored as hashes, which are lines of code unique to that particular section of the sample RNA; the workflow then compares these hashes to the hashes of potential transcripts originating from the reference genome (Bray et al., 2016). After reads have been made, Kallisto samples and resamples RNA read alignments to determine their accuracy in a statistical process called bootstrapping.

Once samples have been quantified, they are not of much use until they can be statistically analyzed and compared to one another. Sleuth is a program used on DNA Subway that has been developed for analysis of RNA-Seq data that has been quantified by Kallisto (Pimentel, 2017). Sleuth utilizes Shiny by RStudio, an open source R package for building web applications, to provide statistical algorithms for investigation of differential expression.

The goal of my research project is to use the DNA Subway interface and RNA-seq approach to investigate and quantify differential gene expression in wild-type and *stm ath1* double mutant inflorescences. Inflorescences include flowers at a wide range of developmental stages, the shoot meristem and stem tissue. I hypothesize that expression

of *HAE* will be reduced in the *stm ath1* double mutant samples compared to wildtype. By using the non-directed approach, I expect to identify a set of differentially expressed genes that can serve as a starting point for other researchers investigating the transcriptional machinery that controls the processes of organ boundary formation and abscission zone differentiation in *Arabidopsis*.

2. MATERIALS AND METHODS

I. Plant materials, growth conditions and genotyping

Arabidopsis seeds were sterilized, planted and grown under the same conditions as previously described (Childers, 2018) by Victoria McClearn and Jill Thiede. The seed stocks used were Landsberg *erecta* (wild-type) and *stm/+ ath1-5.* Since *stm ath1-5* double mutant fruit have multiple developmental defects and are infertile (Malone, 2018; Childers, 2018; Anderson, 2019), seeds were collected from plants homozygous for the *ath1-5* allele and heterozygous for the *stm* allele. The genotyping approaches used for identifying *stm ath1-5* mutant plants were carried out as previously described (Childers, 2018) by Victoria McClearn, Jill Thiede and other members of the Liljegren lab.

II. RNA-Sequencing

RNA Preparation and Quality Analysis

RNA samples were prepared using the RNeasy© Plant Mini Kit (QIAGEN) by Dr. Liljegren, Adam Harris and Jill Thiede. Each sample was prepared from 57-102 mg of inflorescence tissue which included flowers from the earliest stage of development through self-fertilization (stages 1-14; Smyth, 1990). The amount of tissue and number of inflorescences collected for each sample are detailed in Table 1.

The quality of five wild-type and seven *stm ath1* mutant RNA samples was tested using a 2100 Bioanalyzer Instrument (Agilent Technologies Inc., Santa Clara, CA) in the lab of Dr. Scott Baerson (USDA/Natural products). RNA electropherograms are produced by the Bioanalyzer, and RNA integrity numbers (RIN) are assigned on a scale from 1-10, with 10 representing an RNA sample with the least amount of degradation. These values are generated from the electropherograms by taking the ratio of the area under the 18S and 25S rRNA (ribosomal RNA) peaks over the total amount of area under the graph; representative electropherograms are shown in **Figure 6**. Three samples from each genotype with the highest RIN values (**Table 1**) were sent to Cold Spring Harbor Laboratories for RNA-sequencing.

Table 1. Assessment of RIN Samples

Sample Name	**Tissue Weight (mg)**	**Number of Inflorescences**	**RNA Integrity Number (RIN)**
Wild-type A	45.8	6	8.9
Wild-type B	46.1	6	8.2
Wild-type D	**64.4**	**12**	**9.2**
Wild-type X	**56.5**	**4**	**9.4**
Wild-type Y	**65.8**	**4**	**9.0**
stm ath1 M2	96.8	8	8.9
stm ath1 M3	96.5	8	8.8
***stm ath1* M4**	**83.6**	**8**	**9.5**
***stm ath1* M5**	**67.6**	**8**	**9.3**
stm ath1 M6	59.7	8	9.3
***stm ath1* M7**	**101.6**	**10**	**9.5**
stm ath1 M8	64.7	10	8.7

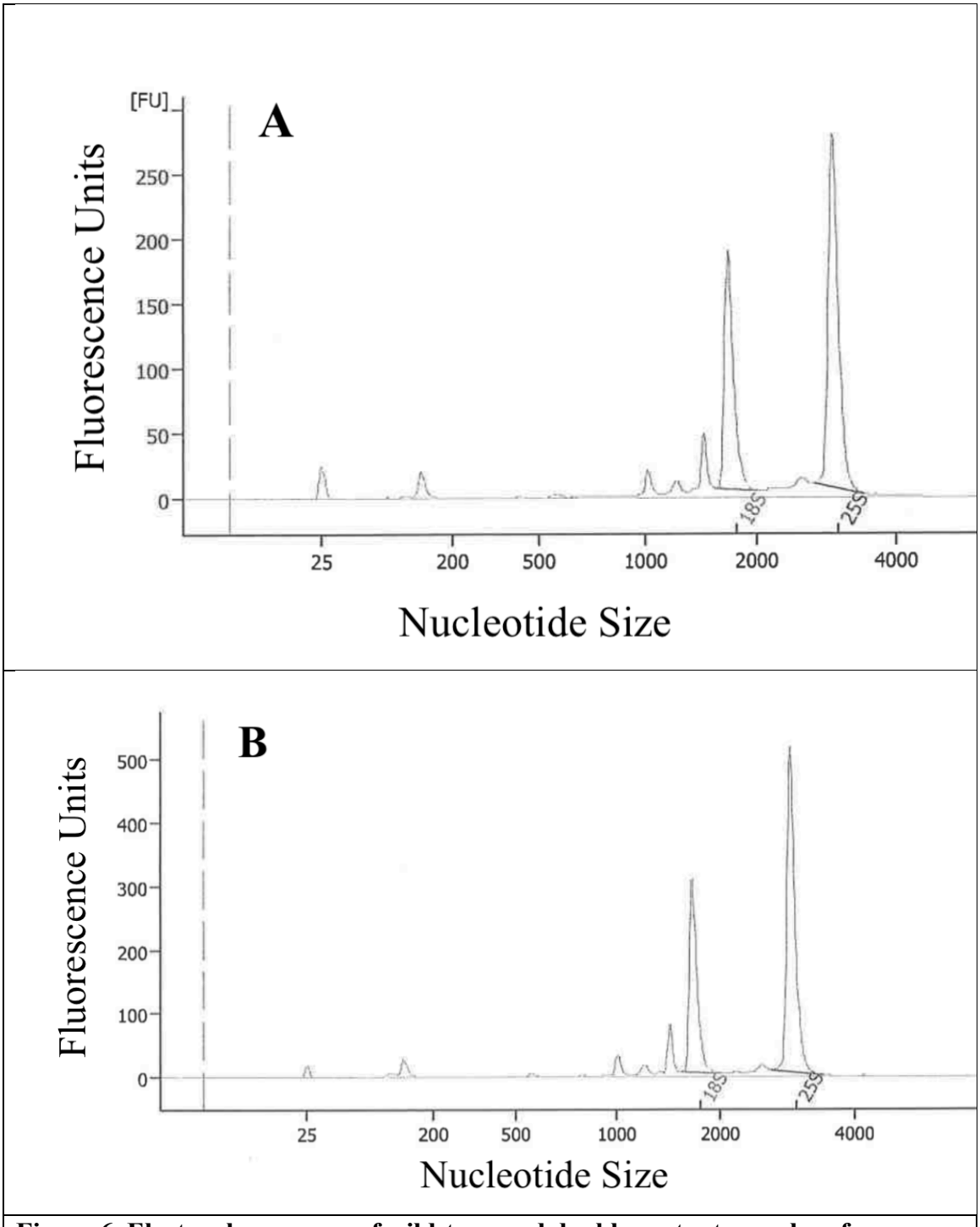

Figure 6. Electropherograms of wild-type and double mutant samples of *Arabidopsis* inflorescence RNA. Arbitrary fluorescence units (FU) are plotted as a function of RNA size in nucleotides (nt).
(A) Electropherogram of wild-type X sample with a RIN value of 9.4.
(B) Electropherogram of *stm ath1* M7 sample with a RIN value of 9.5.

RNA Sequencing

Once purification and a quality check of the RNA samples had been completed by, the samples were sent to Cold Spring Harbor Laboratories (CSHL) for mRNA enrichment, cDNA library construction, and library sequencing. There are several RNA-Seq library protocols that are offered by CSHL including: poly-A selection, ribo-depletion, and size selection (Kukurba and Montgomery 2015). This project focuses on mRNA transcripts that ultimately result in a protein product; therefore, we elected to use poly-A selection useful for sequencing mRNA by selecting for RNA species that contain a poly-A tail (mRNA). Enriched mRNA is then converted to cDNA via reverse transcription with sequencing adaptors ligated onto the ends of each cDNA for differentiation between strands. CSHL currently houses several Illumina Next Generation Sequencers (Illumina, San Diego, CA) for producing raw reads of the cDNA library.

Creation of the Project in DNA Subway's Green Line

The DNA subway Green Line platform was used as a pipeline to analyze the sequencing reads generated by Cold Spring Harbor Laboratories. Samples were read in paired end fashion against the selected organismal genome, *Arabidopsis thaliana*. Paired end reads are more effective than single end reads due to the fact that reads are done in the forward and reverse resulting in increased accuracy and precision. The project creation graphical user interface (GUI) can be seen in **Figure 7A**. Upon project creation, the pipeline GUI is presented and the Kallisto workflow is selected (**Figure 7B**).

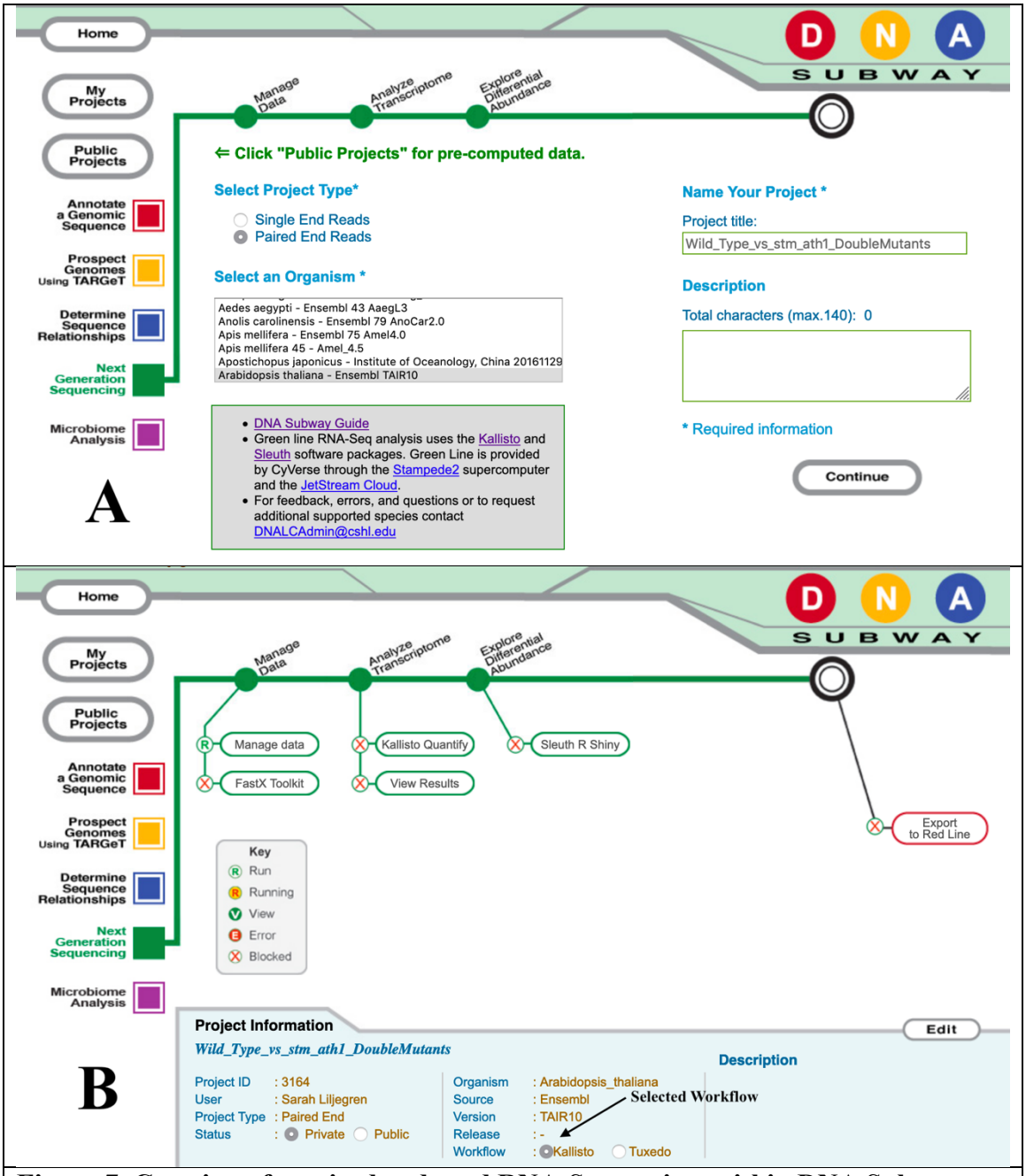

Figure 7. Creation of a paired end read RNA-Seq project within DNA Subway. **(A)** Project creation screen displaying a paired end, *Arabidopsis* project being created. **(B)** DNA Subway's Green Line interface displaying the Kallisto workflow.

Sample Quality Check:

Before data can be read and analyzed by Kallisto, the samples must be checked for quality through the "FastX Toolkit" (**Figure 7B**). The FastX Toolkit utilizes the FastQC program to check samples for per base sequence quality, per base sequence quality scores, per base sequence content, per sequence guanine, cytosine (GC) content, and sequence length and sequence distribution. If one or more of these criteria are not met, the samples will be unable to be analyzed effectively.

Forward and reverse reads were then properly paired using the manage data interface (**Figure 8A**). Following sample pairing, FastQC was ran, and the results were viewed in the FastQC Report screen (**Figure 8B**)

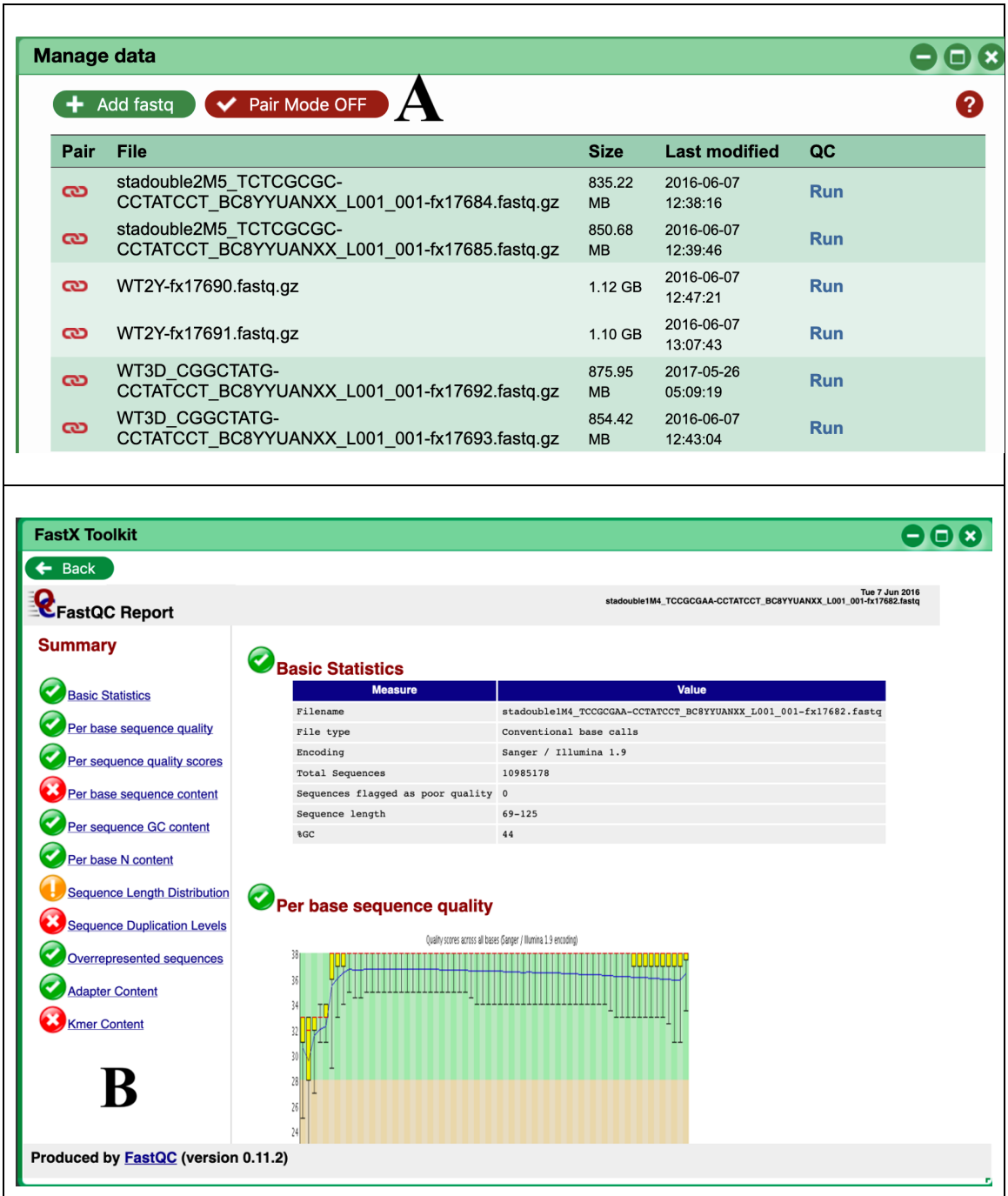

Figure 8. Quality check of samples carried out by FastQC.
(A) Pairing of proper left and right reads of desired samples within the "Manage Data" interface of DNA Subway.
(B) The FastQC Report interface used for quality checking of your submitted samples.

Data Analysis via Kallisto:

Once all samples had been checked for quality, they were quantified using Kallisto. Paired samples were given a sample name based on their replicate number, and a condition based on their genotype. Wild-type samples were given the sample name "WTX" (X denoting the replicate number) and *stm ath1* double mutant samples were given the name "MUTX". The condition of the samples was also based on genotype (either WT for wild-type samples or MUT for double mutant samples (**Figure 9A**).

Once each sample had a designated sample name and condition, the Kallisto program was ran, and the results were available for download (**Figure 9B**). The results of quantification of the RNA transcripts read could then be directly manipulated by downloading or by visualization using Sleuth.

Exploration of Differential Abundance

The Sleuth web application was used to view and display the results from the quantification of reads by Kallisto. Every read made by Kallisto was found in a table, as well as information about their significance values and fold change. This application was used to produce plots and illustrations used for differential expression comparison including principal component analysis (**Figure 11**), heatmap (**Figure 12**), differential expression boxplot of *HAESA* (**Figure 13**), and volcano plot (**Figure 14**).

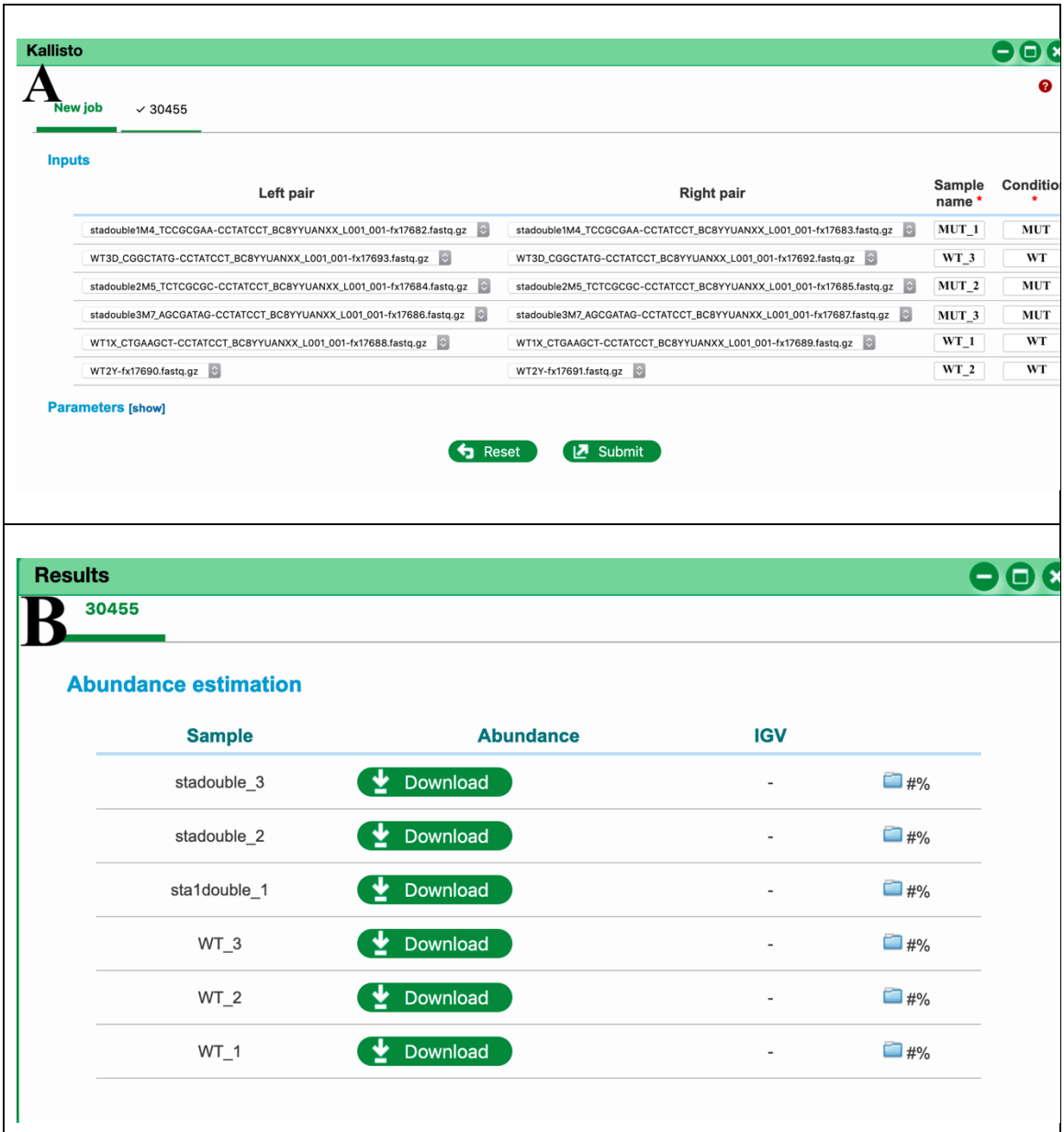

Figure 9. Kallisto input and output interface.
(A) Kallisto input interface where sample names and conditions are applied to samples.
(B) Kallisto output interface where raw reads are available for download.

3. RESULTS

After multiple RNA samples from the inflorescences of wild-type and *stm ath1* double mutant plants were prepared, their quality was tested using a Bioanalyzer (see **Table 1**). Three replicates for each genotype were sent to CSHL for preparation of cDNA libraries and cDNA sequencing. Raw reads were deposited at DNA subway in a private account for the Liljegren lab. I began this project by analyzing the results from quantification and pseudoalignment by the Kallisto package run on CyVerse servers through the DNA Subway interface. Visualization of data was done through the RStudio package "Sleuth".

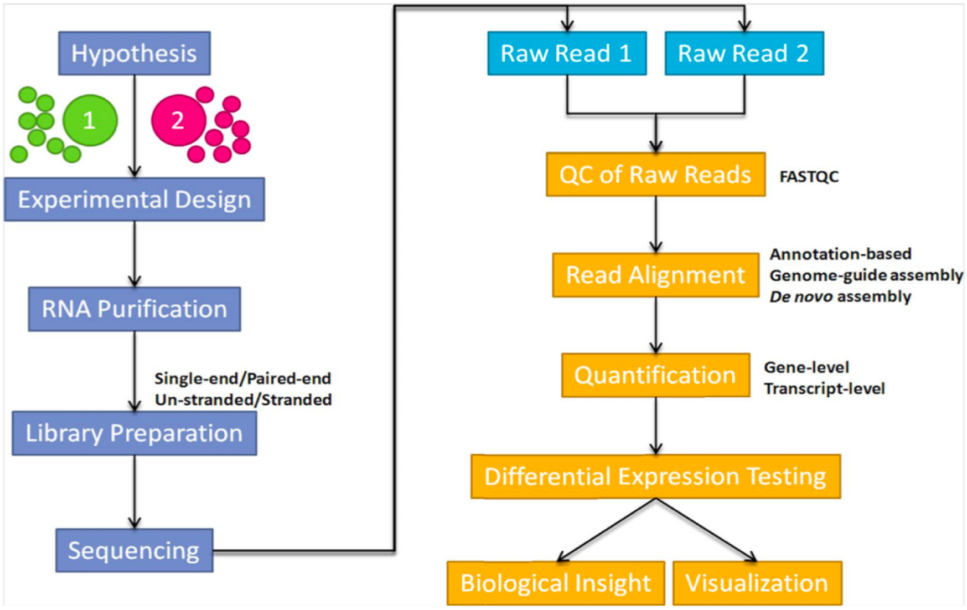

Figure 10. A flowchart detailing the experimental design of RNA-Seq. The left side in blue denotes RNA sampling from two experimental groups in replicate followed by cDNA preparation and Next-Generation Sequencing. Millions of raw reads can then be fed into a bioinformatics pipeline represented in yellow. This provides differential gene analysis between sample groups 1 and 2 shown in green and pink.
(image credit: Enke 2017)

The basic experimental design for RNA-Seq is shown in **Figure 10**. Initially, RNA replicates from two sample groups are prepared for analysis, a control group and an experimental group. In this case, these are our wild-type and *stm ath1* mutant inflorescence samples. RNA is purified from these samples and cDNA libraries are prepared for each independent sample via reverse transcription, and the cDNAs from each library are sequenced. Raw reads, the individual bases of all cDNA within the sample, are then checked for quality and aligned against the reference genome of *Arabidopsis*. The raw reads and quantification data can then be visualized and compared. Visualization of the data provides many advantages including analyzing correlations

between replicate RNA variation, differential expression quantification, and global differential expression analysis.

Determining sample correlation and variation

To begin, sample correlations were considered. Genes in an organism can generate different transcription products depending on the site that transcription starts and on splicing variations. These differences can be detected by RNA-seq. Determining the similarity between samples can be useful in understanding the range of different RNA transcripts produced in the control group and how they are affected in the experimental group.

Principal Component Analysis (PCA) is a statistical procedure that utilizes orthogonal transformation to convert a set of possibly correlated values onto a plot against two principal components (**Figure 11**). PCA analysis is useful for finding hidden patterns in the data. Linear correlations between samples are found and plotted on a two-dimensional plot for easier viewing and comparison analysis. For this pilot study, the *stm ath1* double mutant replicates were found to have a similar negative correlation as displayed by their close grouping (**Figure 11**). Likewise, the wild-type replicates can be seen grouped toward the positive end of the plot denoting similar inter-sample correlation (**Figure 11**). While discovering transcriptional starting points and splicing variants of RNA were not the main goal of this study, this data could prove to be useful to members of the lab for further analysis in the future.

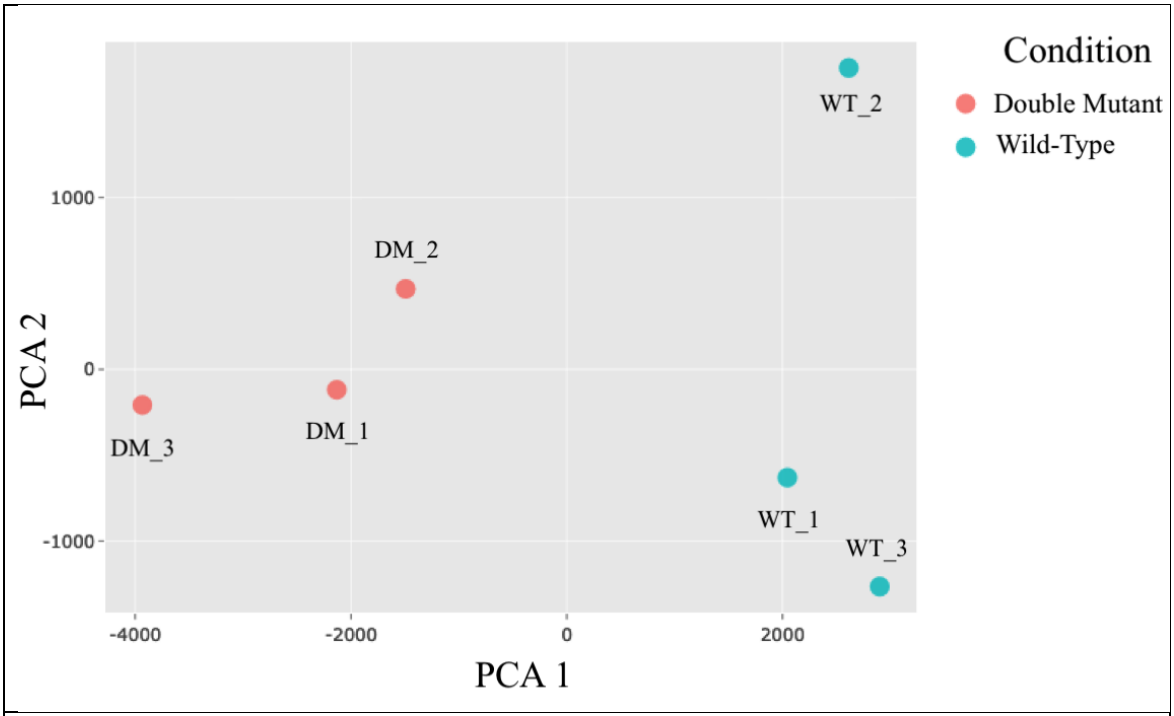

Figure 11. Projection of sample variance against the first two principle components. Linear correlations are plotted two-dimensionally for easy reference and comparison. Genetically similar samples are usually seen grouped close together.

Jean-Shannon Divergence (JSD) is another method used to measure the similarity between two probability distributions. JSD correlations are generally plotted on a heatmap. This provides an easy way to view the similarity between samples in an intuitive manner (**Figure 12**). In contrast to PCA analysis (**Figure 11**), JSD correlation provides direct comparison to other individual samples. Samples that have a similar cDNA library are shown in a lighter color. A comparison score of zero, shown in white, indicates that the two samples are identical. Observation of the heatmap shows the darker blue boxes are between the WT3 and DM1-3 samples, and between the WT2 and DM1-2 samples, which suggests, as expected, that the cDNA populations show more variance between the wild-type and *stm ath1* genotypes. In any case, it is reassuring that they are not strong correlations between the replicates of either genotype.

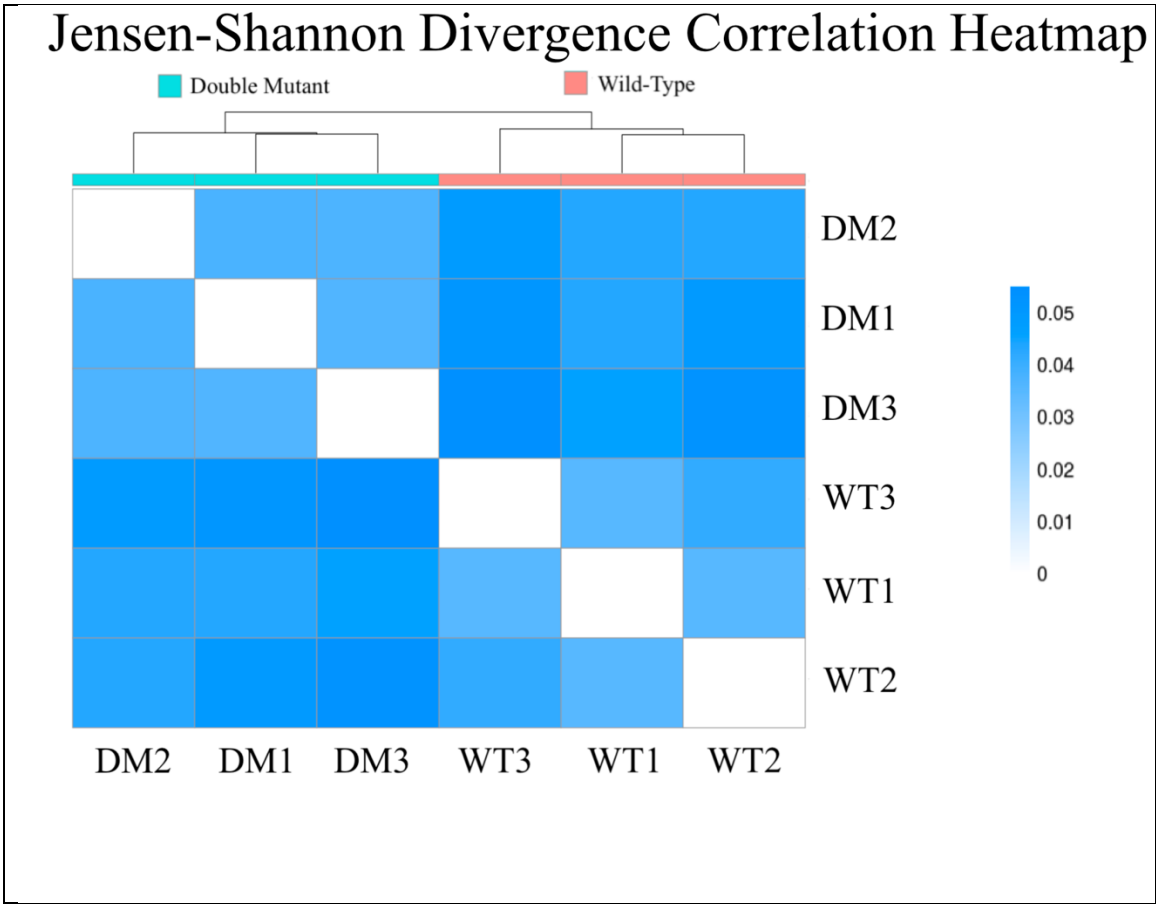

Figure 12. A heatmap displaying the measure of similarity between samples using Jean-Shannon Divergence. Lighter color denotes a higher similarity between sample RNA transcript and variance.

Expression of *HAESA* is altered in *stm ath1* flowers

To display the directed approach of RNA-Seq, the expression levels of the receptor-like kinase gene of interest, *HAESA*, were looked at specifically. The variation and abundance of a gene's RNA transcripts are displayed in boxplots. The expression data are quantified in counts, which are the number of transcript reads within a given sample that overlap with a particular genomic locus. Each of the transcript reads are then counted and the information is stored within a text file outputted by the package. In conjunction with Sleuth, these reads are then able to be easily compared to one another in graphics with the ability analyze any gene that was read in the initial sample. A

differential expression boxplot for my gene of interest, *HAESA*, is displayed in **Figure 13**. The results show that there is reduced expression of *HAE* in *stm ath1* double mutants when compared to wild-type plants. This evidence concurs with previous research using a molecular marker for *HAE* expression (**Figure 4**; Raybourn, 2016).

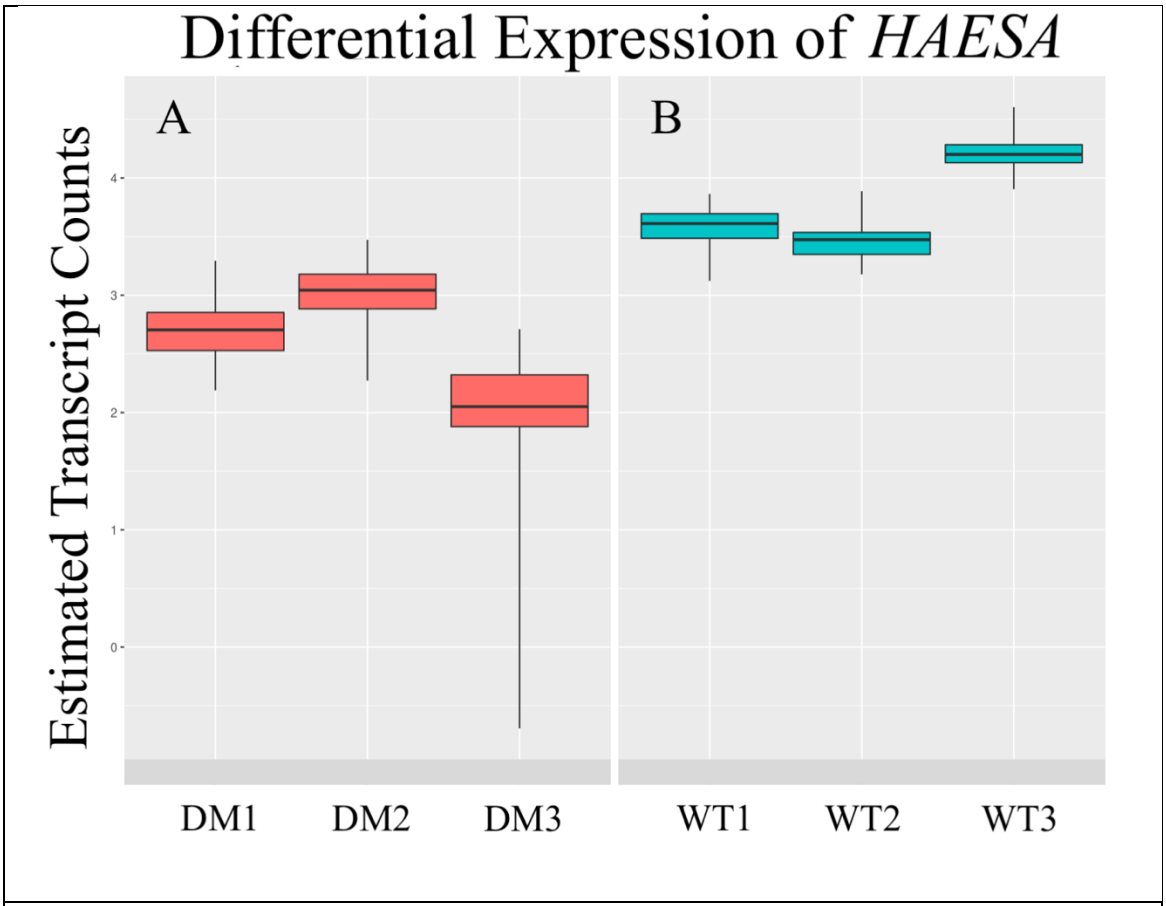

Figure 13. Boxplot comparing *HAESA* expression in wild type and *stm ath1* mutant RNA transcriptomes.
(A) The estimated transcript counts of *stm ath1* double mutant samples that contain overlapping segments of the *HAESA* gene.
(B) The estimated transcript counts of wild-type samples that contain overlapping segments of the *HAESA* gene.

Exploration of global differential gene expression

Another tool available through RNA-Seq is the process of solo transcriptome profiling. This allows the researcher to see the expression of genes on a global scale. Using this information, I was able to locate genes that are similarly downregulated to the gene *HAE*. Genes that are have similar regulation to the target gene can give hints to the molecular pathway involved in organ abscission, and allow for location of other genes that are affected by the mutation of *ATH1* and *STM* homeobox genes.

Changes in expression are quantified using a *b*-value. The *b*-value is the ratio of the expression of a gene in the wild-type samples over the expression of the same gene in the *stm ath1* double mutant samples. A positive fold change suggests that the gene is expressed at lower levels in the mutant compared to wild-type. The significance of each read is quantified by the *q*-value. The *q*-value represents the *p*-value, or probability value, adjusted for false discovery rate. A *p*-value normally denotes the probability that all tests will produce a false positive, while a *q*-value states that only the probability that significant tests will result in a false positive. Scientific experiments typically require a *p*-value of 0.05 to be considered significant. However, with *q*-value because you eliminate insignificant tests, the accepted value is 0.1. Since the -log10(0.1) is one, values that display a *q*-value greater than one are considered significant.

Genes are mapped to a volcano plot for easy viewing (**Figure 14)**. The x-axis of the graph displays the fold change (*b*-value) ratio of wildtype gene expression vs. double mutant expression. The y-axis displays the -log10 of the *p*-value adjusted for false discovery rate (*q*-value) for easier interpretation and graphical representation. Significantly upregulated and downregulated genes are normally located toward the

upper right and left portions of the plot, respectively. Reads with significant *q*-values (greater than one) are shown as pink dots, and those that did not return a significant *q*-value are shown in gray as false (**Figure 14**).

A list of 24 genes with a *b*-value greater than two and a -log10(*q*-value) greater than one are shown in **Table 5**. Gene loci are shown using TAIR nomenclature. Following "AT" *for Arabidopsis thaliana* is the number that denotes which of the five chromosomes the gene is located on. The five-digit gene code is found after the "G". Splice variants of gene transcripts are denoted by a decimal point and the assigned splice variant number. Genes are displayed in order of fold change (*b*-value) from greatest to least. Representative Gene Ontology (GO) Biological Process and Molecular Function terms were selected from TAIR for genes with sets of assigned terms. Gene Ontology is a consortium whose goal is to "provided an up-to-date, comprehensive, and computational model of biological systems" by assigning molecular function and biological process to genes in short terms for categorical purposes (Ashburner et al., 2000). Categorization of massive gene libraries allows for researches to have quick and easy access to genes of interest. This list of genes can provide our lab as well as others with future genes of interest that could hold value in the genetic and molecular mechanisms of abscission.

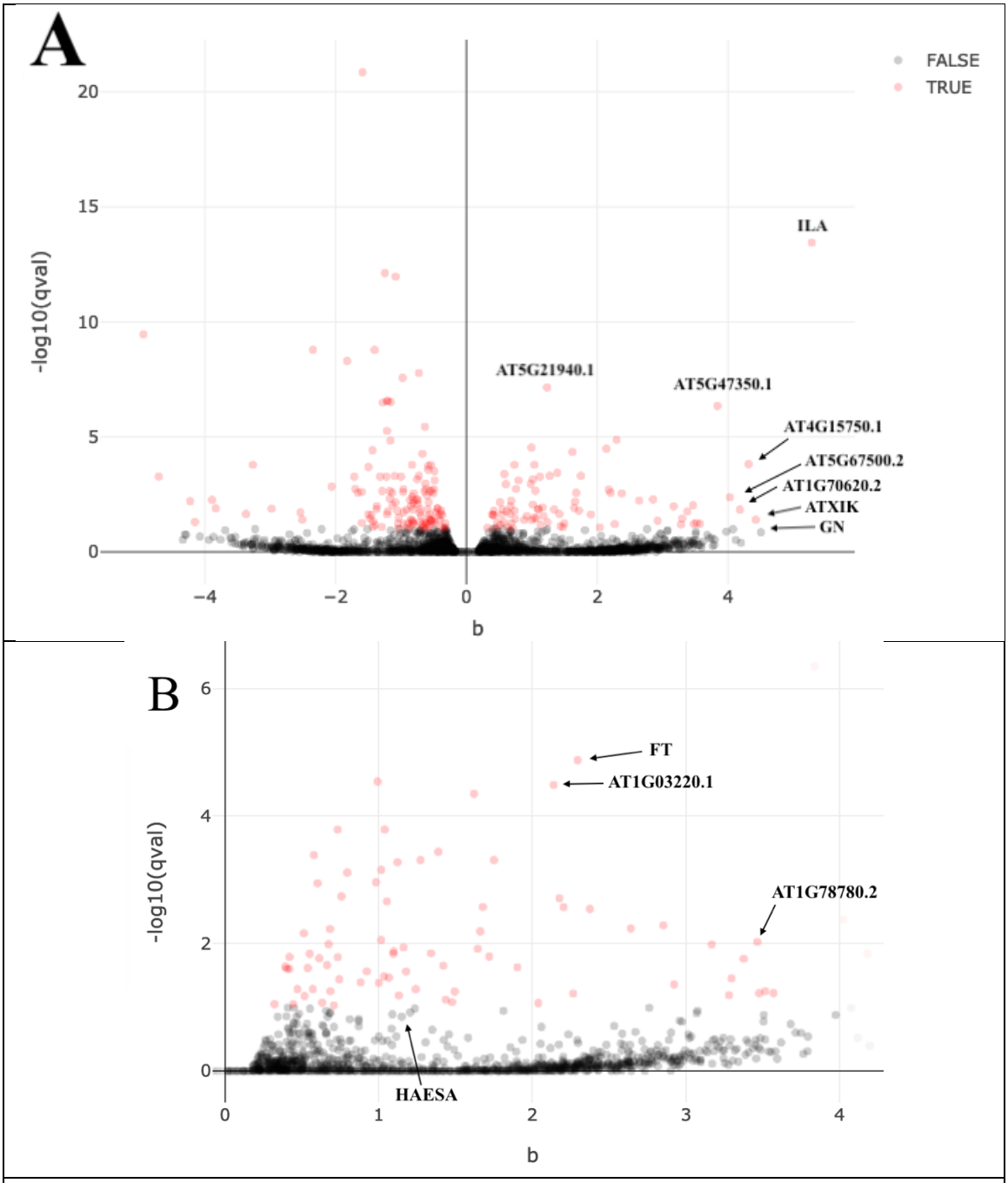

Figure 14. Differential expression between *stm ath1* and wild-type inflorescences **A)** Genes with a positive and negative fold change (*b*-value) are depicted on the right and left sides of the y-axis, respectively. A positive fold change indicates an increase in gene expression (transcript reads) in the experimental group (*stm ath1* mutant) compared to the control group (wild-type). A negative fold change indicates a decrease in gene expression in the experimental group compared to the control group. The y-axis denotes the significance value, which is calculated as –log10(*q*-value). Significant values are shown in pink, and insignificant values (less than one) in gray. A subset of genes that display a large fold change and significant *q*-value are labeled. **B)** Enlarged view of the genes with a positive fold change.

Table 2. *Arabidopsis* genes with reduced expression in *stm ath1* inflorescences compared to wildtype.

Gene Name	Locus	-log10(*q*-value)	*b*-value	GO Biological Process	GO Molecular Function
ILA	AT1G64790.2	13.45	5.28	Defense response	-
ATXIK	AT5G20490.3	1.40	4.42	Actin filament movement	Motor activity
-	AT4G15750.1	3.82	4.31	Negative regulation of catalytic activity	Enzyme inhibitor activity
-	AT1G70620.2	1.84	4.18	-	-
-	AT1G56100.2	2.37	4.03	-	-
-	AT5G47350.1	6.35	3.84	-	Palmitoyl hydrolase activity
SEN1	AT4G35770.2	1.22	3.57	Aging; Response to jasmonic acid	-
-	AT1G0950.2	1.25	3.52	Lignin biosynthetic process	Alcohol dehydrog-enase activity
EXPA25	AT5G39300.1	1.22	3.48	Cell wall loosening and organization	-
-	AT1G78780.2	2.02	3.46	-	-
SCPL43	AT2G12480.3	1.76	3.38	Proteolysis	Serine-type carboxype-ptidase activity
DFR	AT5G42800.1	1.45	3.30	Anthocyani n-containing compound biosynthetic process; Redox process	Coenzyme binding; Flavanone 4-reductase activity

Table 2 cont. *Arabidopsis* genes with reduced expression in *stm ath1* inflorescences compared to wildtype.

Gene Name	Locus	-log10(*q*-value)	*b*-value	GO Biological Process	GO Molecular Function
SEEDSTK	AT4G09960.1	1.19	3.28	Ovule and carpel development	DNA binding
BAN	AT1G61720.1	1.98	3.17	Flavonoid biosynthetic process	Anthocyanidin reductase activity; Coenzyme binding
KTI1	AT1G73260.1	1.36	2.92	Defense response to bacterium and molecule of fungal origin;	Endopeptidase inhibitor activity
AAS	AT2G20340.1	2.28	2.85	Response to wounding	Carboxylase activity
TPS18	AT3G14520.1	2.24	2.64	Sesquiterpene biosynthetic process	Cyclase activity; Terpene synthase activity
SWEET9	AT2G39060	2.54	2.37	Carbohydrate transport; Nectar secrtion	Protein binding; Sugar transmembrane transporter activity
FT	AT1G65480.1	4.88	2.3	Cell differentiation	Protein binding
-	AT3G49270.1	1.21	2.27	-	-
SBT3.14	AT4G21630.1	2.57	2.20	Induced systemic resistance; Proteolysis	Serine-type endopeptidase activity
GILT	AT4G12960.1	2.71	2.18	-	Catalytic activity

Table 2 cont. *Arabidopsis* genes with reduced expression in *stm ath1* inflorescences compared to wildtype.

Gene Name	Locus	-log10(*q*-value)	*b*-value	GO Biological Process	GO Molecular Function
-	AT1G03220.1	4.49	2.14	Protein catabolic process	Aspartic-type endopepti-dase activity
SAG12	AT5G45890.1	1.07	2.04	Aging; Leaf senescence	Cysteine-type endopepti-dase activity
HAESA	**AT4G28490**	**0.85**	**1.15**	**Floral abscission**	**Kinase**

4. DISCUSSION

Using RNA-Seq technology, I was able to quantity the differential expression of *HAE* in wild-type and *stm ath1* double mutant inflorescences. My results showed that there is a diminished total count of *HAE* RNA transcripts in the *stm ath1* double mutant compared to wild-type (**Figure 13**). These results are consistent with a previous study using a *HAE:GUS* marker (**Figure 4**; Raybourn, 2016). However, the significance value of *HAE* in this pilot study displayed a -log10(*q*-value) of 0.85, which is in the range of a false discovery, and the fold change of 1.15 is modest (**Table 2**).

A goal of the transcriptome analysis I conducted was to use a non-directed approach to look for candidate genes that may be regulated by the STM and ATH1 transcription factors (see volcano plot in **Figure 14**). For this pilot study, I focused on identifying genes that were expressed two-fold less in *stm ath1* inflorescences compared to wild-type and displayed a significant value higher than one. Twenty-four genes were found; some of which may play roles in either establishing organ boundaries or promoting abscission zone differentiation.

This evidence could warrant investigation of these genes individually to determine the effects that disruption of *STM* and *ATH1* have on their expression. Reverse transcriptase quantitative PCR (qRT-PCR) and RNA in situ hybridization are two methods that can be used to directly analyze gene expression (Udvardi et al., 2008, Fransz et al., 1996). RT-qPCR is used to further qualitatively detect gene expression by creating complementary DNA transcripts from RNA. RT-qPCR uses reverse transcriptase to reverse transcribe RNA into its DNA complement similar to RNA-Seq, but expression

levels are quantified in real time using florescence arrays. RT-qPCR can detect very low levels of RNA because the cDNA complement is very stable and is thus useful in qualitatively looking at the most minimal gene expression. RNA in situ hybridization is used to quantitatively measure a specific RNA sequence in a section of tissue (in situ) using a fluorescent reporter probe.

Single-cell RNA-Seq (scRNA-Seq) is a recently developed technology that improves the resolution of transcriptome profiling down to that of a single cell of interest. Cells can be isolated using a variety of methods such as: limiting dilution, micromanipulation, flow-activated sorting, laser capture microdissection, and microfluidics (Kolodziejczyk et al., 2015; Hwang et al., 2018). Once cells have been isolated, a number of different scRNA-Seq technologies can then be implemented. These protocols differ somewhat in their workflow, but most of them follow similar qualitative and quantitative analysis to bulk RNA-Seq that was performed in this study.

In conclusion, my studies suggest that further, more refined studies of differential expression in *stm sth1* double mutants is warranted. Subsequent research should be done to analyze the organ boundary regions of *Arabidopsis* flowers rather than entire inflorescences. Furthermore, flowers can be quantitatively compared at specific stages of development to show the progression of differential expression in these boundary regions. Another worthwhile direction is to study the transcriptome profiles of the *stm* and *ath1* single mutants compared to the *stm ath1* double mutant. Identifying genes that STM and ATH1 regulate independently and jointly would give a more complete picture of how these transcription factors function to establish floral organ boundaries.

BIBLIOGRAPHY:

Aida, M., and M. Tasaka. (2006) Morphogenesis and patterning at the organ boundaries in the higher plant shoot apex. *Plant Molecular Biology*, **60**: 915–928., doi:10.1007/s11103-005-2760-7.

Anderson, K. (2019) Analysis of septum defects in Arabidopsis organ boundary mutants. Department of Biology, The University of Mississippi.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**: 796–815., doi:10.1038/35048692.

Arnaud, N., and V. Pautot. (2014) Ring the BELL and tie the KNOX: roles for TALEs in gynoecium development. *Frontiers in Plant Science*, **5**., doi:10.3389/fpls.2014.00093.

Ashburner, M., et al. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**: 25–29., doi:10.1038/75556.

Bray, N., Pimentel H., Melsted, P., and Pachter L. (2016) Near-optimal probabilistic RNA-Seq quantification. *Nature Biotechnology*, **34**: 525–527., doi:10.1038/nbt.3519.

Breuil-Broyer, S., Morel, P., de Almeida-Engler, J., Coustham, V., Trehin C. (2004) High-resolution boundary analysis during Arabidopsis thaliana flower development. *The Plant Journal*, **38**: 182–192., doi:10.1111/j.1365-313x.2004.02026.x.

Bürglin, R., and Affolter, M. (2015) "Homeodomain Proteins: an Update. *Chromosoma*, **125**: 497–521., doi:10.1007/s00412-015-0543-8.

Childers, K. (2018) *Fruit Structure in Arabidopsis Thaliana Organ Boundary Mutants.* Department of Biology, The University of Mississippi.

Cho, S. K., Larue, C., Chevalier, D., Wang, H., Jinn T., Zhang, S., and Walker, J. (2008) Regulation of floral organ abscission in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, **125**: 15629–15634., doi:10.1073/pnas.0805539105.

Chu, Yongjun, and David R. Corey. (2012) RNA Sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, **22**: 271–274., doi:10.1089/nat.2012.0367.

Cole, M., Nolte, C., Werr, W. (2006) Nuclear import of the transcription factor SHOOT MERISTEMLESS depends on heterodimerization with BLH proteins expressed in discrete sub-domains of the shoot apical meristem of Arabidopsis thaliana. *Nucleic Acids Research*, **34**: 1281–1292., doi:10.1093/nar/gkl016.

Conesa, Ana, et al. (2016) A survey of best practices for RNA-Seq data analysis. *Genome Biology*, **17**., doi:10.1186/s13059-016-0881-8.

Dewey, C. (2014). Lecture for BMI 877: Measuring transcriptomes with RNA-Seq [PowerPoint Slides]. Retrieved from Slideplayer site.

Enke, R.A. (2017). Lecture for Bio 481: FASTQC analysis and HISAT alignments using CyVerse (part 2) [Handout]. Retrieved from Bepress site.

Fransz, Paul F., et al. (1996) High-resolution physical mapping in Arabidopsis thaliana and tomato by fluorescence in situ hybridization to extended DNA fibres. *The Plant Journal*, **9**: 421–430., doi:10.1046/j.1365-313x.1996.09030421.x.

Gubert, C. M., Christy, M. E., Ward, D. L., Groner, W. D., and Liljegren, S. J. (2014) ASYMMETRIC LEAVES1 regulates abscission zone placement in Arabidopsis flowers. *BMC Plant Biology*, **14**., doi:10.1186/s12870-014-0195-5.

Gómez-Mena, C., and R. Sablowski. (2008) ARABIDOPSIS THALIANA HOMEOBOX GENE1 establishes the basal boundaries of shoot organs and controls stem growth. *The Plant Cell Online*, **20**: 2059–2072., doi:10.1105/tpc.108.059188.

Hwang, B., Lee, J. H., and Bang, D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, **50**., doi:10.1038/s12276-018-0071-8.

Jinn, T. L., Stone, J. M., and J. C. Walker. (1999) HAESA, an Arabidopsis leucine-rich repeat receptor kinase, controls floral organ abscission. *National Center for Biotechnology Information*. U.S. National Library of Medicine, **9**.

Kolodziejczyk, A. A., Kim, J. K., Scensson, V., Marioni, J. C., and Teichmann, S. A. (2015) The technology and biology of single-cell RNA sequencing. *Molecular Cell*, **58**: 610–620., doi:10.1016/j.molcel.2015.04.005.

Kukurba, K. R., and S. B. Montgomery. (2015) RNA sequencing and analysis. *Cold Spring Harbor Protocols*, **2015**., doi:10.1101/pdb.top084970.

Leslie, M. E., Lewis, M. W., Youn, J. Y., Daniels, M. J., and Liljegren S. J. (2010) The EVERSHED receptor-like kinase modulates floral organ shedding in Arabidopsis. *Developmental Biology*, **331**: 409., doi:10.1016/j.ydbio.2009.05.084.

Long, A., Moan, E., Medford, J., and Barton, M.K. A member of the KNOTTED class of homeodomain proteins encoded by the STM gene of Arabidopsis. *Nature*, **379**: 66–69., doi:10.1038/379066a0.

Malone, H. (2018) Characterizing the effects of mutations in STM and ATH1 on floral organ development in Arabidopsis thaliana. Department of Biology, The University of Mississippi.

Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D., and Koornneef, M. (1998) Arabidopsis thaliana: a model plant for genome analysis. *Science*, **282**: 662–682., doi:10.1126/science.282.5389.662.

Merchant et al. (2016) The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLOS Biology* doi:10.1371/journal.pbio.1002342.

Mukherjee, K., Brocchieri L., and Burglin, T. R. (2009) A comprehensive classification and evolutionary analysis of plant homeobox genes. *Molecular Biology and Evolution*, **26**: 2775–2794., doi:10.1093/molbev/msp201.

Ori, N., Eshed, Y., Chuck, G., Bowman, J.L., and S. Hake (2000) Mechanisms that control *knox* gene expression in the *Arabidopsis* shoot. *Development*, **127**: 5523–5532.

Palmer, S. (2018) Quantifying abscission defects in mutant Arabidopsis flowers. Department of Biology, The University of Mississippi.

Pimentel et al. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods* 14, 687-690.

Raybourn, D. (2016) Investigation of the expression of the HAESA receptor-like kinase as regulated by the STM and ATH1 homeodomain transcription factors in Arabidopsis thaliana**.** Department of Biology, The University of Mississippi.

Rubenstein, B., and A. C. Leopold. (1964) The nature of leaf abscission. *The Quarterly Review of Biology*, **39**: 356–372., doi:10.1086/404326.

Rutjens, B., Bao, D., van-Eck-Stouten, E., Brand, M., Smeekens, S., and Proveniers, M. (2009) Shoot apical meristem function in Arabidopsis requires the combined activities of three BEL1-like homeodomain proteins. *The Plant Journal*, **58**: 641–654., doi:10.1111/j.1365-313x.2009.03809.x.

Smyth, D. R., Bowman, J. L., and Meyerowitz, E. M. Early flower development in Arabidopsis. *The Plant Cell Online*, **2**: 755–767., doi:10.1105/tpc.2.8.755.

Udvardi, M. K., Czechowski, T., Scheible, W. R. (2008) Eleven golden rules of quantitative RT-PCR. *The Plant Cell Online*, **20**: 1736–1737., doi:10.1105/tpc.108.061143.

Yu, H., and Huang, T. Molecular mechanisms of floral boundary formation in Arabidopsis. *International Journal of Molecular Sciences*, **17**: 317., doi:10.3390/ijms17030317.