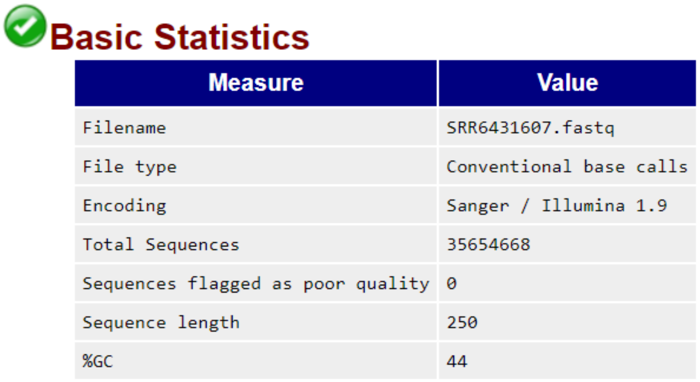
ABE516 Project 2
MSD1 regulates pedicellate spikelet fertility in sorghum through the jasmonic acid pathway
Chen Chen and Ken Youens-Clark

Introduction

This 2018 study used a RNA-seq analysis to identify the transcription factor *MSD1* in sorghum that leads to doubling and tripling of grain number per panicle (GNP) due to the activation of jasmonic acid (JA) biosynthesis and signaling during the development of the flowers. As sorghum is an agronomically important feed crop, understanding the mechanism that increases seed production will lead to better yields from a crop which is also more resistant to drought conditions and grows on less fertile soil than other grasses such as wheat and barley. Importantly, the study found that the mutant gene can be incorporated and expressed into other sorghum genetic backgrounds, perhaps opening the path to cross the mutant into standard growth lines.

The authors "used whole-genome sequencing bulk-based segregant analysis to identify the casual gene." The "genomic DNA from 50 homozygous *msd1-1* mutants selected from an F_2 population were pooled and sequenced to about 27X coverage." "Whole-genome sequencing was also performed on the parental line BTx623 and an additional 18 independent homozygous *msd* mutants, including *mds1-1* and *msd1-2*." "Quantitative reverse transcriptase PCR (qRT-PCR) of various tissues revealed that *MSD1* expression was enriched during inflorescence development." "To characterize the temporal and spatial expression patter of *MSD1*, [the authors] performed RNA *in situ* hybridization." "To interrogate targets of *MSD1*, [the authors] performed transcriptome profiling by RNA-seq on developing panicles from four development stages of WT [wild-type] and *msd1-1* with three biological replicates." The authors found that stage 4 contained "[t]he number of the genes most differentially expressed." The data is available in the NCBI SRA under the accession "SRP127741."

Data Preparation



✓ Basic Statistics

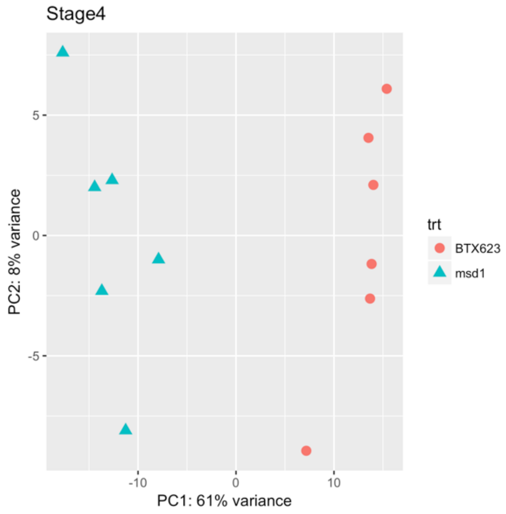
Measure	Value
Filename	SRR6431607.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35654668
Sequences flagged as poor quality	0
Sequence length	250
%GC	44

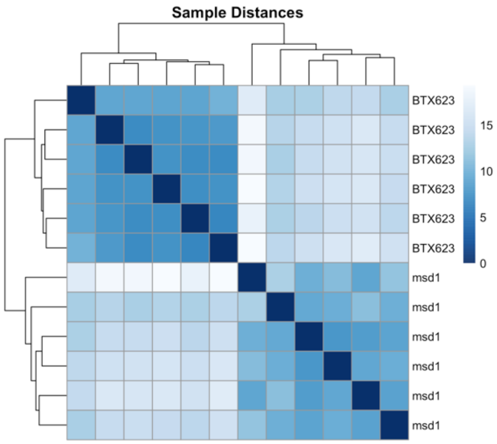
To analyze the data, we downloaded the 36 RNA-seq samples from the project page from the NCBI Sequence Read Archive (SRA) using the "fastq-dump" tool provided by the "sratoolkit." As it was unnecessary to analyze the entire data set, we first downloaded just 100K and 500K sequences for each sample but found there was insufficient data for our analysis. At 1M sequences, we began to find enough signal, but then decided to analyze only on the 12 "stage_4" samples, 6 of the WT (BTx623) and 6 of the mutant (msd1), as the authors reported this stage had the greatest number of differentially expressed genes. For this analysis, we used 10M sequences for each of the 12 samples.

Since the authors used the STAR aligner, we decided to do likewise. This required us to download a reference sorghum genome which we found on the Gramene.org FTP site[1] as well as a Gene Transfer Format (GTF) file[2] that defines the sorghum gene structures. We then ran STAR with the "--runMode genomeGenerate" command to create a genome index against which we aligned each of the 36 samples resulting in Sequence Alignment Map (SAM) files. These were converted to binary (BAM) format using the "samtools view" command distilling about 12MB of alignment data which we needed to begin analysis in the R language.

Sample Visualization

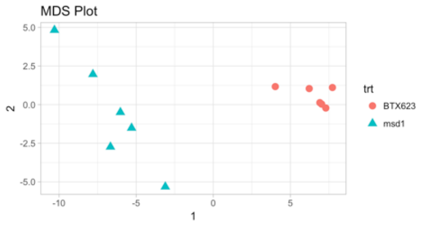
After log-normalizing the data, we used a principal component analysis (PCA) plot to show that there is, indeed, first principal component in the gene expression at this stage accounting for a total of 61% of the variance.





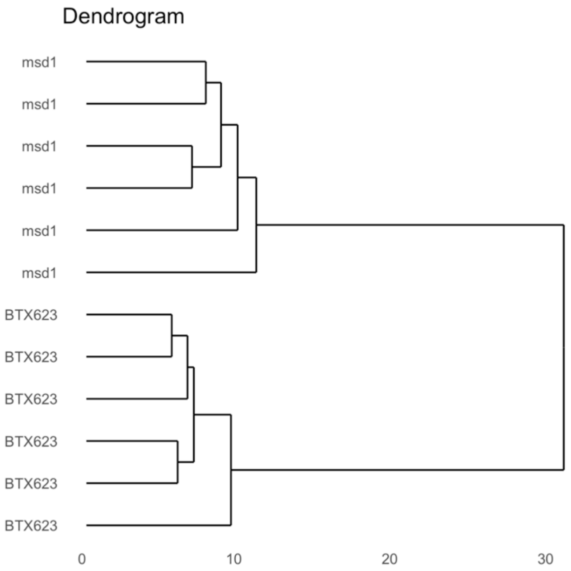
We can create a similarity matrix we used to cluster the samples to see that they group as expected on the WT and *msd1* genotypes.

Another method plot that is very similar to the PCA can be produced with the multi-dimensional scaling (MDS) function in R. It also shows the expected clustering of samples.



[1] ftp://ftp.gramene.org/pub/gramene/release-56/fasta/sorghum_bicolor/dna/Sorghum_bicolor.Sorghum_bicolor_NCBIv3.dna.toplevel.fa.gz

[2] ftp://ftp.ensemblgenomes.org/pub/release-38/plants/gtf/sorghum_bicolor/Sorghum_bicolor.Sorghum_bicolor_NCBIv3.38.gtf.gz



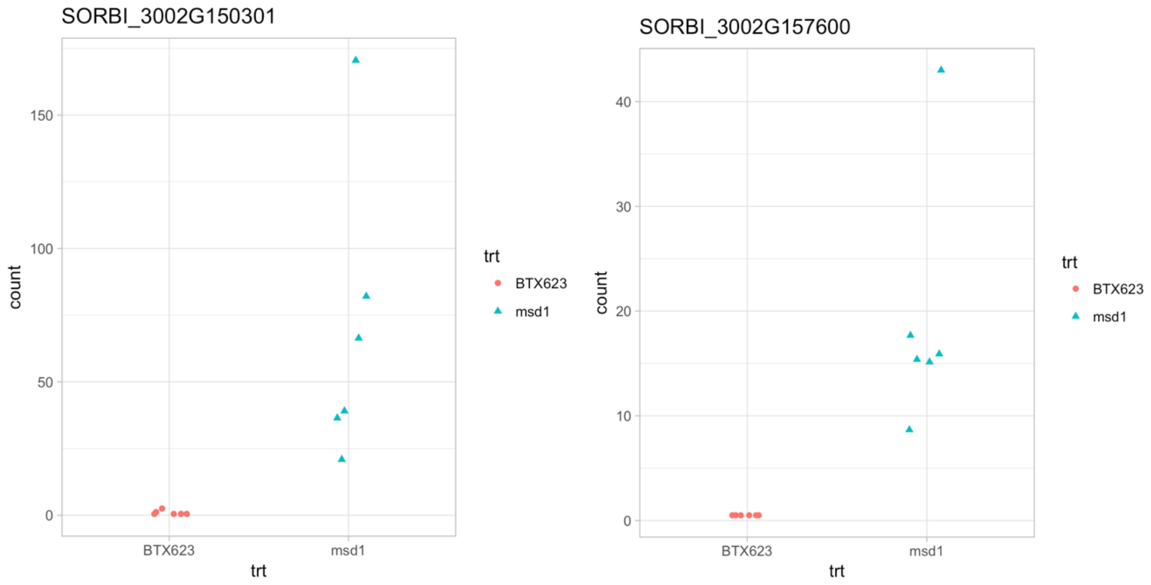
By subtracting the similarity matrix from 1, we created a distance matrix which can be used to generate a dendrogram which further supports the clustering of the samples by genotype.

Differentially Expressed Analysis

We used DESeq2 to do differential expression analysis with FDR=0.05. Here we detected 133 differentially expressed genes shown in the table below.

With 556 non-zero read counts	
LFC>0 (up-regulated)	73 (13%)
LFC<0 (down-regulated)	60 (11%)

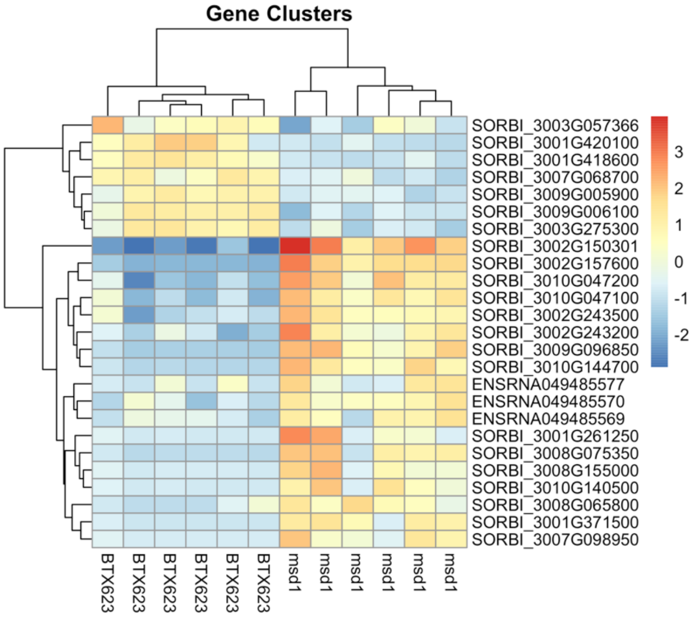
Looking more closely at these genes, we can examine how their expression varies in the 12 plants. Here are the top two for an example.



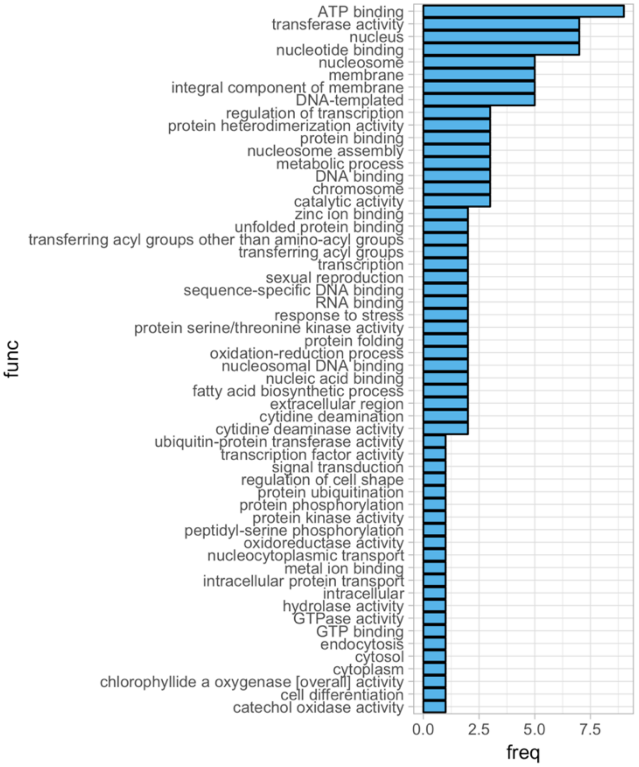
Clustering Analysis

We can also cluster samples based on their most expressed genes, and again we find the sample distinctly cluster in the same way. For space we will only use the top 25 most differentially expressed genes.

This heatmap clearly illustrated that first 7 genes are highly expressed in BTX623 (wild type) and the last 18 genes are highly expressed in *msd1* (mutant). Then we want to see the functions of these genes.



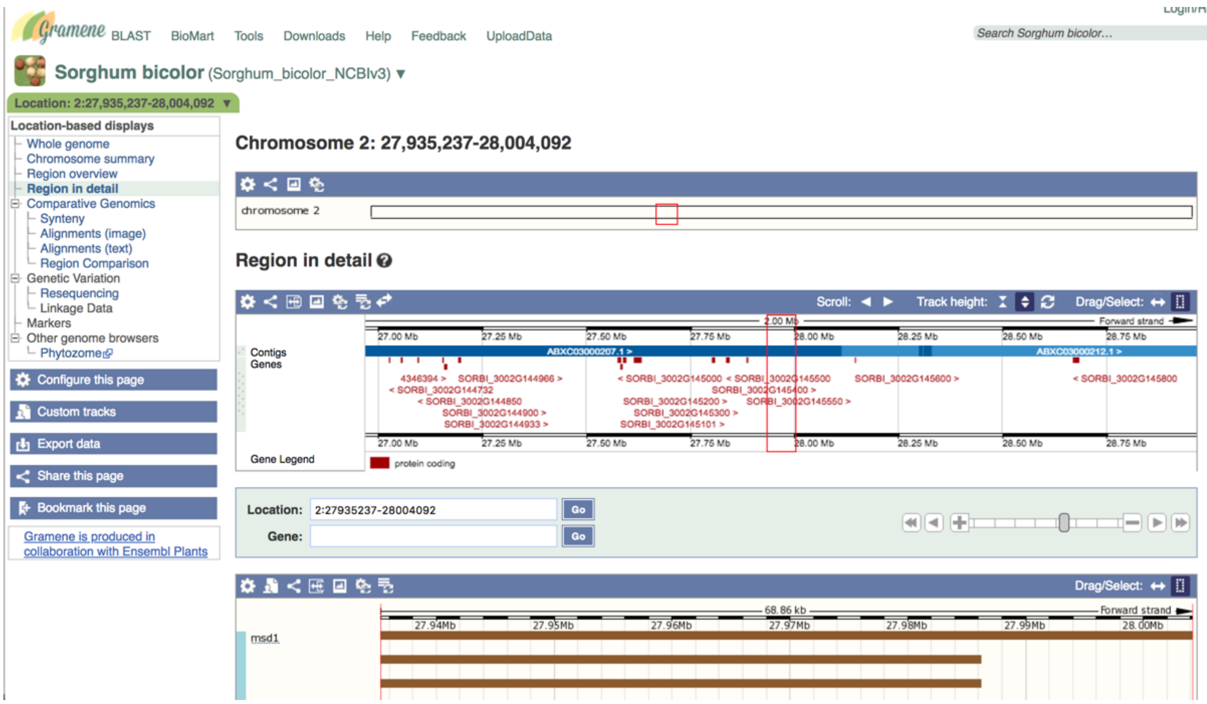
Gene Ontology



If we query for all the genes which were differentially expressed with a log-fold change greater than 1 and with a p-value < 0.05, we find 48 (remembering that only used a subset of the original data). As the R library "AnnotationHub" has no annotations for *Sorghum*, we used the Gramene.org BioMart tool to query for Gene Ontology (GO) annotations for these genes in order to find out what these genes are doing in the plants. We created a summary of the GO terms most represented. Given the subset of data we processed, we were unable to recapitulate the paper's GO analysis. Also, they used an online service (AgriGO) that was incompatible with our gene IDs.

Viewing Top Genes

To view the genes in the context of the Sorghum genome, we converted the BAM files to BED format using the "bamToBed" program provided by the "bedtools" package. We then upload the BED features to the Gramene.org *Sorghum bicolor* genome browser. Below is a sample view of some of the features on chromosome 2.



Location: 2:27,935,237-28,004,092

Chromosome 2: 27,935,237-28,004,092

Region in detail

Location: 2:27935237-28004092 Go

Gene: Go

Conclusion

Using the first 10 million sequences for each of our 12 samples, we detected 133 differentially expressed genes between stage-4 BTX623 and *msd1* samples. Among them, 73 genes are up-regulated, and 60 are down-regulated. With clustering methods and gene ontologies, we found some interesting patterns. Clustering of samples is very consistent with the background information. Further analysis like pathway analysis and network analysis is needed, and biological validation is recommended.

Reference

Yinping et al. *MSD1 regulates pedicellate spikelet fertility in sorghum through the jasmonic acid pathway.* Nat Commun. 2018 Feb 26;9(1):822. doi: 10.1038/s41467-018-03238-4.