Hi, I’m Pranav, and I’m going to be a junior at UNL this upcoming fall majoring in Computer Science. This summer, I worked under Dr. Chi Zhang to conduct research to identify differentially expressed genes between multiple genetically altered lines of Soybean plants.

**First, let’s dive into the Background:**

The main goal of this experiment was to isolate genes that increase the oil content in soybean seeds as soybean oil is a potential source of biofuel.

Plants capture carbon dioxide to synthesize macromolecules such as carbohydrates and lipids. In soybeans, possessing an efficient pathway for this is essential to increase oil content in its seeds. While we know that the links exist between pathways and oil production, this experiment hopes to understand relations between specific genes which is key to develop plants with better yield.

The experiment contains 4 lines of soybean plants each with 3 replicas as shown in Table 1— the experimental ones being plants that have the WR1, the WR1 + DGAT, and the WR1 + KasII gene treatments edited in their genome. It’s worth noting that WR1 is an enabler protein that increases lipid generation hence why it is also being tested with the other genes.

Each of these plants have their gene count measured using a high-throughput RNA-sequencing tool at three distinct time points R5, R5/6, R6—as shown here which are stages of seed development. Gene counts represent the number of times a certain gene is present in the genome in parts per million. These have varying impact on the plants as for instance, a large count of an oil production gene means that more oil is produced. Our goal is to find genes with counts that have large differences between genotypes.

Each sample of obtained data follows a strict naming convention such as (B3\_2) where ‘B’ represents the time period, the 3 represents the genetic treatment, and the 2 represents the replica number.

**Second, let’s look at the methods:**

The raw counts obtained from the experiment are shown in Figure 1. We, however, cannot compare these for two reasons

1. The gene counts for each sample could have been acquired from a different part of the plant thus unbalancing the whole sample itself. For example, the gene reads between a leaf and a stem are obviously going to be very different as their tissues are different. AND
2. The data is large so any small p-value would still lead to huge amounts of false positives. For instance, if we are experimenting with 50000 genes and a p value of 0.01 which is conventionally very gooood, the result still has 500 false positives which is not great.

To combat this, we used Bioconductor’s DESEQ2 package which uses logarithmic normalization methods to compare the genes. Logarithmic methods are better because it reduces variability in data. Figure 2 does a great job explaining this.

Here we compare the log of all the gene counts of three samples, A1\_1, A1\_3, and B2\_4. The more clustered the graph is, the more similar the two samples are. So, it makes sense that A1\_1 & A1\_3 are closer than A1\_1 & B2\_4 as the former comparison is between replicas. If we compared any sample to itself, it would be a perfect line across the diagonal.

In addition to the logarithmic normalization, DESEQ2 also utilizes a design formula as shown in Figure 3. Here, it controls for the time values across samples while analyzing the differentially expressed genes among the different genotypes. Each of these comparisons have 2 groups of samples—one control and one experimental and their designations are shown in Table 2. We’ve isolated the samples so that in each comparison there is only 1 independent variable we are testing for. DESEQ2 also uses an adjusted-p value calculation through the Benjamini and Hochberg method that is way more accurate than the regular one thus reducing the number of false positives.

**Third, let’s discuss the results:**

The results from running the data through the DESEQ2 model are shown in Figure 4. For the purposes of this presentation, we will discuss comparison between wild type and WR1 although the same methods were applied on all of the treatments. There are two important variables that we will focus on for each gene. They are log2foldchange and adjusted p value.

Log2foldchange is the difference between the log of the gene counts and tells us how different two genes are. If the log2foldchange for gene “x” is 4, that means that the WR1 samples have 2^4 time as many reads of that gene as the Wild type samples. The adjusted p value is calculated internally so that it more accurate than the original p-value thus limiting false positives.

We have sorted the results from Figure 4 to only show genes with a log2FoldChange > 0.7 AND p-adj < 0.005. That way we can assure that the most significantly different genes and ones with highest probability of being true positives are only considered.

We then used soybase (an online soybean genomics tool) to analyze these significant genes to find their biological function. A pie chart of this is displayed in Figure 5. The 3 biggest parts of the graph are carbohydrate/lipid synthesis and transport which corroborates the idea that WR1 is responsible for increased oil production and maintenance among other functions.

**To conclude,** differential gene analysis in soybeans can help us find the intricate relations that each gene has on the overall plant and on each other. The results that we have displayed show us the way each gene impacts several others. Another aspect of differential analysis we can explore is how the genes change over time and how it varies across genotypes. This can help us construct better plant lines that have an increased oil yield.

Finally, I would like to acknowledge the Nebraska Center for Energy Sciences Research for directly supporting this research. I would also like to thank Dr. Chi Zhang for mentoring and guiding me through this project.

Thank you everyone for listening to this presentation and I hope you have a great day!