Abstract

The purpose of this study was to better understand mechanisms of the effect of hydrogen peroxide on *Beta vulgaris* seed germination. We hypothesized that genetic expression rather than morphological changes results in efficient seed germination, and will align with previous phenotypic studies of the effects of hydrogen peroxide (H_2O_2) on the germination. The goal was to identify candidate gene sequences in seeds treated with H_2O_2 that phenotypically expressed early and consistent germination. As there is no fully annotated genome for the *Beta vulgaris* species, it was necessary to adapt genomes from other plants such as the well-documented *Arabidopsis thaliana*, otherwise known as thale cress, among others. Using other species can yield varied annotations of gene sequences. In order to mitigate this error, multiple annotated gene databases were used. In our investigation, using the GenBank database and others, several candidate genes were found with slightly varying annotation "hits" for each database. A general consensus can be formed as to which of the annotations are of the most relevance. To further delineate the mechanisms, additional investigations using biochemical technique such as PCR (polymerase chain reaction) or qPCR (quantitative real-time polymerase chain reaction) will be needed.

Executive Summary

As the sugar beet is very important for global sugar production, optimization of the crop efficiency is very beneficial for production. Experiments conducted in our lab have shown that when sugar beet seeds are soaked in a solution containing 0.3% hydrogen peroxide, they germinate significantly more rapidly and consistently. This trait is applicable to commercial sugar production, so it is important to more thoroughly study the mechanisms. Firstly, it would be useful to identify if genetic mechanisms are involved in triggering the H_2O_2 treated seeds to germinate more efficiently. Secondly, it will be useful to sequence the genetic code so it can be used to develop genetically modified seeds in the future. Bioinformatic tools are a useful tool to conduct preliminary investigations to identify if genetic alterations are responsible for the changes seen in the H_2O_2 treated seeds, and not just a morphological change in the seeds. In these experiments, I used the bioinformatic tools such as GENBANK to look for differences in gene sequences between the seeds soaked in distilled water and seeds soaked in 0.3% H_2O_2 solution. Through this analysis, I was able to identify several sequences of DNA that likely contribute to the improved germination. As predicted, higher expression levels of these gene sequences were identified. Further genetic analysis using biochemical techniques can hone in the specific gene mutations that confer phenotypic properties to the H_2O_2 treated seeds that are advantageous. Bioinformatic tools provide a very efficient way to conduct preliminary studies to identify the genetic mechanism(s) involved, and in the future developing genetically modified seeds.

A Differential Expression Analysis of Effect of Hydrogen Peroxide on *Beta vulgaris* Seed Germination

Vinay Hiremath* and Dr. Mitchell McGrath**

**International Academy, Bloomfield Hills MI*

***Research Geneticist and Adjunct Professor at USDA Agricultural Research Service and Michigan State University, East Lansing MI*

Project Confirmation Number: 0008589

Introduction

The sugar beet, accounting for an estimated 30% of the worlds' sucrose supply, is an immensely important crop that necessitates proper investigation (Fairfood International). In the past, efforts have been made to increase the success rates of germination of the sugar beet, *Beta vulgaris*. It has been found through preliminary experiments that the seeds of the species *Beta vulgaris* germinate significantly more successfully when placed in a solution of 0.3% H_2O_2 as opposed to a solution of distilled H_2O. In the H_2O_2 solution, there is a noticeable difference in both germination time, which is shorter, and the percentage of seeds germinated, which is higher, after four days. For commercial production, which the sugar beet is important for, both of these aspects are greatly beneficial to crop yield because of the rapid and consistent germination. After obtaining DNA sequences of the seeds in both the control and experimental group, I will attempt to use various bioinformatic tools to investigate the specific DNA sequences that are only expressed in the experimental H_2O_2 group.

This work is important because of its potential benefits for commercial crop harvests. Since direct H_2O_2 treatment in the field cannot be used as it will later damage the crop, changes at the genetic level must be investigated to identify pathways that control this change. It is likely that the change in success rates is a direct effect of a change in the genes being expressed by these seeds.

Hypothesis

Our hypothesis is that the H_2O_2 treated seeds which germinate rapidly and consistently are a result of changes at the genetic level caused by treatment with H_2O_2, unlike the distilled water treated controls. The newly developed properties are not just a result of a change in morphological characteristics in the H_2O_2 treated seeds.

Methods

I used a large suite of bioinformatics tools, all of which are open-sourced under the GNU General Public License and are therefore freely redistributable. In order to first build an index for the sequence files so that they can be more easily indexed, I first needed to generate detailed quality reports, so I used the FASTQC toolkit. This suite handles the FASTQ format, which is a text-based sequence for storing sequence data. It is the successor to the FASTA format, and its main difference is that it incorporates quality scores in the Phred33 format (Trapnell et al.).

These sequences are read in 100bp sets that are in no particular order, which is why alignment onto a genome of known sequences for *Beta vulgaris* is necessary, which is shown in Figure 1. Furthermore, sets of these sequences often have overlapping sections, which by definition can be detected by an identical sequence in these reads. This is demonstrated in Figure 2. As contigs need to be placed in the proper position on the genome, areas between these aligned contigs known as scaffolds are displayed in Figure 3.

For this investigation, I first used a tool called Bowtie which allows strong compatibility with the assembly program I used, known as Tophat. Using this suite of programs, I was able to assemble the data I had gathered onto the RefBeet 0.9 genome from Max Planck Institute for Molecular Genetics. These alignments are reported in BAM files, a binary compressed version of the SAM format, together making up the predominant output formats of next-generation sequence alignment tools (Trapnell et al.). With this, I was able to analyze the remaining assembled data sets using the Cufflinks suite. Cufflinks generated transcript indexes of each data set for both H_2O and H_2O_2 of each seed variety. Furthermore, Cuffdiff was able to find the sections of each contig that was different between the experimental and control data sets while avoiding those similar between the seed variety data

sets, which helped to eliminate extraneous results. Following this, I deemed it helpful to focus the data so that it could be more easily managed and analyzed further. To gather more consistent data, I found identical sections between the two files. This, in other words, is a file containing the locations of the shared differences between H_2O and H_2O_2 for both seed types (USH20 and MSR3), which is an improvement over the original difference files as it finds only those differentially expressed in both seed varieties. To convert these contigs to actual sequences as represented in the RefBeet-0.9 genome, I used a custom script that extracts the base sequences from the reference genome by applying the contig number and position given by Cuffdiff to the genome. Annotated genome data will be collected from sources using the Basic Local Alignment Search Tool (BLAST) through databases such as the Arabidopsis Information Resource (TAIR), among others. The BLAST system finds "regions of local similarity between protein or nucleotide sequences," as it calculates the statistical similarities of the matches (Bergman et al.). Queries to the BLAST are character strings of nucleotide or amino acid codes, often represented with a descriptive line in the FASTA format, which is further explained below. The alignment score is found by "assigning a value to each aligned pair of letters" and then adding these values along the alignment (Bergman et al.). The annotated databases output predicted the roles of bases in each different scaffold between the control and experimental data sets. Lastly, with the Cufflinks output, I was able to generate detailed expression plots such as volcano, scatter, and box plots using a tool called CummeRbund. This tool parses the Cufflinks output into R objects, which are more suitable for data analysis because of the wide variety of R packages available for this purpose.

The workflow diagram (Figure 4) visualizes the steps of the data analysis used in this investigation.

FASTQ and FASTA Formats

This example is from the MSR3 data set soaked in water. As described, the FASTQ format bundles quality scores, which are displayed on the **fourth line**. The **first line** is a systematic identifier created by the Illumina sequencing system, while the **second line** is the raw sequenced bases and the **third line** is an optional description line. Excerpts of two sample sequences are shown here.

```
@HWI-ST957:100:D0V52ACXX:5:1101:1226:2111 1:N:0:CAGATC
GTGGGCATGAAGTGTGGGGATAGCATGGACTCGCCAGTTGTATCGGCGGTCG...
+
@DFDDAHFDDAEAEHIICCCEHHDIJEHJJIJJHGFHHIIJGGIIIF###############...

@HWI-ST957:100:D0V52ACXX:5:1101:1219:2188 1:N:0:CAGATC
GTGAGCATACCTGTCGGGACCCGAAAGATGGTGAACTATGCCTGAGCGGGGC...
+
@BBADBDEFHHHHGEEHIGGIJHCDBGFGGDGGGGGFGHBFGHCGGHGID@55...
```

This is when compared to the FASTA format, of which an example is shown below. This format only contains an identification line before the sequence line, so it is a two-line system as opposed to the FASTQ four-line system.

```
>scaffold00001 length=4956840
ATCAATGTATGCTTGTTAAATTCTGTTTGTAGCACGACTCGAAACTCGACTTGA...
```

Results

Preliminary Study

The data in Figure 5 was found in a preliminary study I used to test the effect of hydrogen peroxide on seed germination consistency for three *Beta vulgaris* seed varieties. All data is out of a total of 25 seeds for each variety and treatment, and the hydrogen peroxide solution used was at a 0.3% concentration, which is similar to the treatment of the USH20 and MSR3 sequenced seeds. As is clear from the data, the H_2O_2 solution has a clear positive effect on the seed germination consistency.

Sequence Quality Information

Figure 6 represents the Phred33 quality scores output from the Illumina Genome sequencing system. It is created using the FASTQC toolkit, which handles quality analysis for the FASTQ format. The graph shows the quality along each 100bp read of the full sequence, as the sequencing machines that were used handle 100bp reads. This representation is ideal to identify a read error in the machine, because a dramatic drop in quality in one particular position of the read indicates the same problem in each run of the machine. Only the graph for the USH20 set in the H_2O solution is shown for demonstration, but tests for all sets were conducted.

USH20 seed type in H_2O solution	
Total sequences	28485306
Sequence length	100bp
GC content	44%

Figure 7 displays information along each read, with the x-axis displaying base position from 0bp to 100bp. The graph shows the distribution of each base—adenine,

guanine, cytosine, and thymine—along the read. These graphs are only shown for the USH20 seed type soaked in H_2O, simply to demonstrate the types of quality analysis that were performed on these sets. In actuality, the analysis was performed on each data set.

Genome Processing File Sizes

Genome Name	RefBeet 0.9
File Size	575MB
Bowtie2 Index Size	719MB

Cuffdiff Output Information

A box plot as shown in Figure 8 generated using CummeRbund for the Cuffdiff output for the USH20 seed type is shown.

Shared Differences

By combining the two Cuffdiff output files, I ended up with sequences containing sections of 2 contigs and 42 scaffolds as referenced in the RefBeet-0.9 genome. After this file was generated using several command line tools to parse the data, I extracted the sequences from the genome, and the FASTA file containing these 44 subsequences was 164KB in size. This data is significantly smaller as it contains An example location line from this file and its accompanying sequence FASTA sequence lines are shown below.

contig109244:204-849

contig109244:204-849
CCAAAACCAAAATACAAAACTACTTACGT...

Discussion

Example of BLAST Database

Through analysis of the BLAST output annotations, I noticed several hits for *peroxidase*. Fourteen hits were found for *peroxidase* in the TAIR database and another 27 hits in the GenBank database to make a total of 41 instances in the annotations. As this is traditionally an enzyme that helps to break down hydrogen peroxide, I felt that its appearance in a water and hydrogen peroxide differential expression analysis would be relevant. To find the actual relative expression values for this scaffold in the Cuffdiff output, I found the line matching the same scaffold that contained these hits. This line is shown below, with the bolded sections being the most relevant. These values are defined in the following table.

XLOC_008434 XLOC_008434 - **scaffold00062:1143405-1146203** H2O H2O2 OK **170.825 1.38039 -6.95129** 4.96476 6.87866e-07 0.000389594 **yes**

Scaffold number and position	scaffold00062:1143405-1146203
H_2O relative expression	170.825
H_2O_2 relative expression	1.38039
Fold value	-6.95129
Significance	yes

Because the expression value for H_2O is significantly more than that for H_2O_2, it is clear that Cuffdiff reports a higher expression in the H_2O set. However, this data raised my suspicions, as an enzyme such as *peroxidase* that breaks down hydrogen peroxide would be expressed in the H_2O_2-treated seeds. To test the possibility that the treatments were reversed during data processing, I used a known gene (BvGer165) that has been biochemically shown to be highly expressed in H_2O-treated seeds. The following expression results (Figure 9) were

found for this gene in seeds with various treatments in an experiment conducted in the lab I worked at using the northern blot analysis (de los Reyes and McGrath).

Two BLAST hits for this gene were found in a certain scaffold, and the relevant Cuffdiff output values are again shown below. As shown in the table above, Cuffdiff reports a higher relative expression for the H_2O_2-treated seeds, and the difference is significant.

Scaffold number and position	scaffold00152:586250-587620
H2O relative expression	14.8927
H2O2 relative expression	587.037
Fold value	5.30077
Significance	yes

In Figure 9, it is clear that the BvGer165 gene should be expressed significantly more in an H_2O treatment, rather than H_2O_2 as reported by Cuffdiff. By testing this known gene with the relevant Cuffdiff output, it is safe to conclude that the data sets were reversed, possibly due to an orthographic error in the data processing. Now, I identified the scaffolds by classifying them as either signaling, transcription, or expressed gene sequences by using a key BLAST hit within each sequence. An excerpt of the results are shown in the next table, where:

s: signaling
t: transcription
g: expressed gene

Sequence	Example Function	Role
scaffold00001:1203697-1204087	zinc finger	s
scaffold00006:3204-13492	ribosomal protein	s
scaffold00007:281641-282593	nucleolin putative	s
scaffold00010:2235270-2238580	protein phosphatase	s
scaffold00010:2370225-2371363	cytochrome P450 family protein	g
scaffold00010:2384523-2389680	similar to zinc finger protein	t
scaffold00030:491747-493622	myb family transcription	t
scaffold00035:878140-879345	CBL-interacting protein kinase	s
scaffold00054:330364-334342	myb family transcription	t
scaffold00062:1143405-1146203	peroxidase	g

When properly applied, the gene sequences that are found to be expressed by *Beta vulgaris* seeds treated with H_2O_2 can be used to simulate similar growth in seeds that have not been soaked in this solution. The importance of this result lies in the fact that an H_2O_2 solution, although helpful for seed germination, harms further plant development. As a result, it is important to isolate the potential genetic alterations expressed in the H2O2 treated seeds. With this knowledge, using gene modification techniques, the seeds can be manipulated at the genetic level without exposing them to the toxic effects of H2O2. This application has the potential to revolutionize sugar beet growth in a variety of environments that require rapid and reliable seed germination.

Previous reports have not provided a bioinformatic-related study that investigates differential genetic expression to this level, instead opting for studies of the effect of hydrogen peroxide on various aspects of *Beta vulgaris* germination, including its effects on differentiating cultivars of the species. For example, the northern blot protocol experiment completed in the lab I worked at focused on studying expression of several genes in many solutions, but did not attempt to identify the differential expression in DNA that caused these changes. Furthermore, in the study by Dr. Naegele and Dr. McGrath in the Plant Breeding and Genetics Program at Michigan State University, it was found that the "initial conditions that a germinating seed encounters, and its ability to deal with them, affect the rate at which germination occurs" (Naegele and McGrath, 2011). In their study, the "speed and number" of seeds that germinated was found to be different for a non-stressed H_2O condition and a stressed H_2O_2 condition. This study similarly found that H_2O_2 treatments affect seed germination rate and consistency, but did not delve into the genetic expression as done in this investigation. In addition, in a similar study by Dr. McGrath, Dr. Derrico, and others, it was found that "hydrogen peroxide gave a boost to germination" and "germination in water

showed a wide range of germination values," which is very similar to the results I arrived at in my genetic differential expression study (McGrath et al.).

Conclusions and Future Work

The work accomplishes a great deal toward the general study of differential expression of the effects of hydrogen peroxide on *Beta vulgaris* seeds. The most relevant past studies have supported the basic premise that a H_2O_2 treatment aids sugarbeet seed germination, as discussed above. Alternative conclusions are possible but highly unlikely, as the data is supported by both qualitative experiments of the effect of hydrogen peroxide on the seeds themselves, and quantitative experiments by studying differential expression using bioinformatic tools in this investigation.

Although one or a few specific gene sequences have not been identified, with bioinformatic tools alone it is unlikely to arrive at this extent of detail. Maintaining the scope of this project includes not involving a biochemical investigation using reactions such as (PCR) polymerase chain reaction to find exact sequences through amplification of pieces of DNA. However, those steps have been greatly aided by this study, which has found several candidate genes that have a correlation with hydrogen peroxide in *Beta vulgaris*. That said, if more time at a sufficiently equipped lab was available, it would be very beneficial to extend the scope of the project and attempt PCR or qPCR (quantitative real-time polymerase chain reaction) to more accurately focus on a few genes that have a significant effect on germination quality. This will then be more useful to those who wish to increase crop yield of their *Beta vulgaris* seed, as specific experiments may be conducted to try and reproduce this germination quality without the use of H_2O_2 and the damage that it can later wreak.

Illustrations

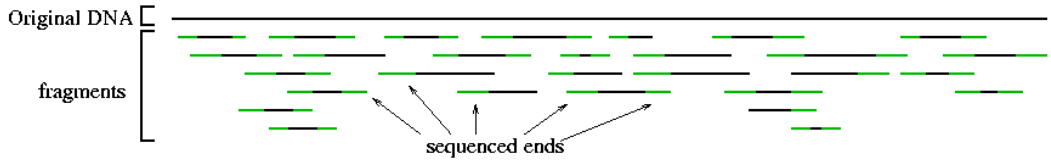


Figure 1: Diagram of sequenced DNA fragments

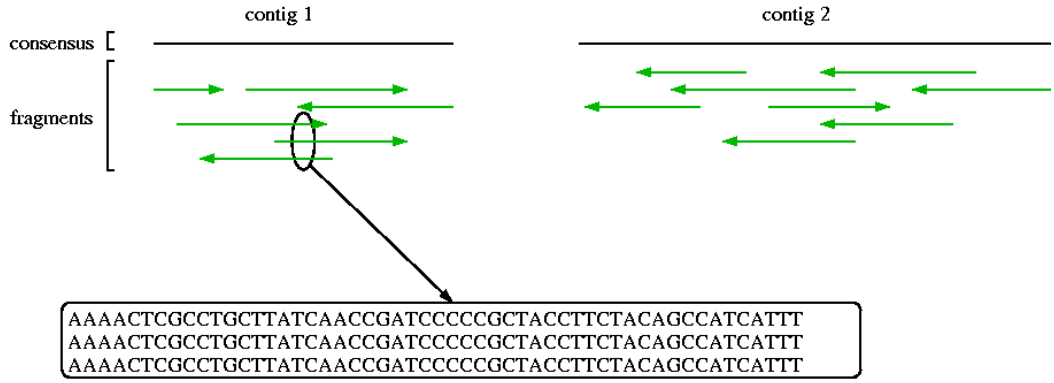


Figure 2: Diagram of contigs formed by duplicate sequences

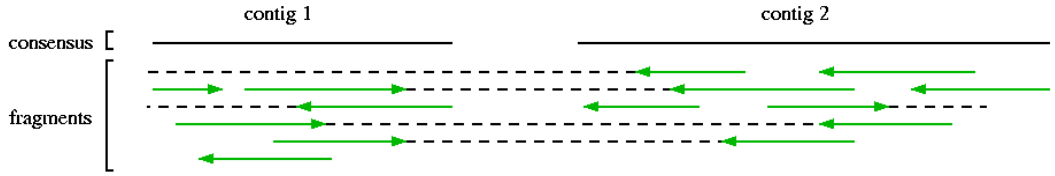


Figure 3: Diagram of scaffolds formed by areas between aligned fragments

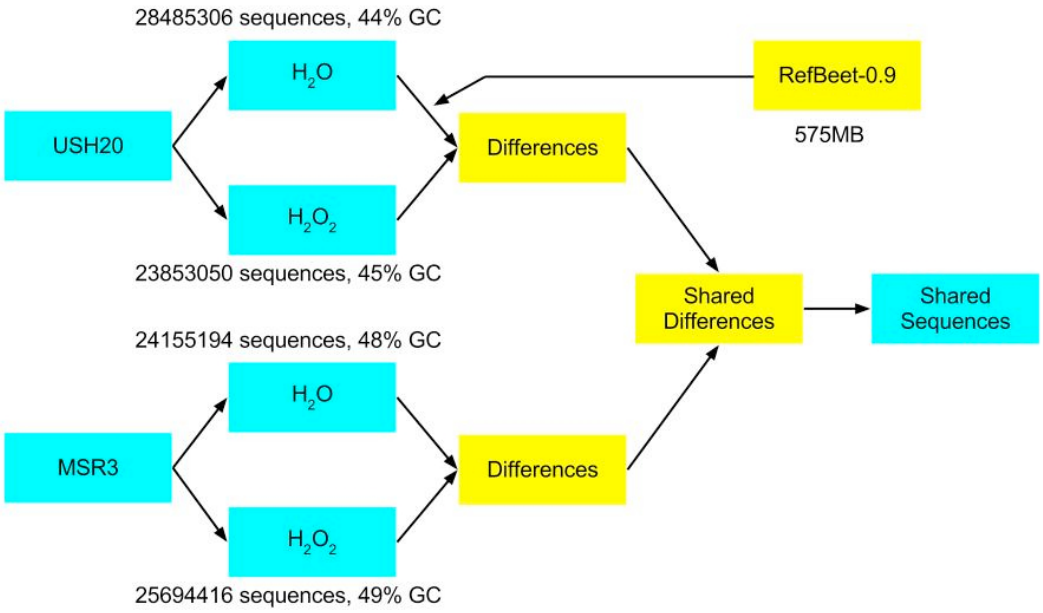


Figure 4: Workflow of sequence analysis

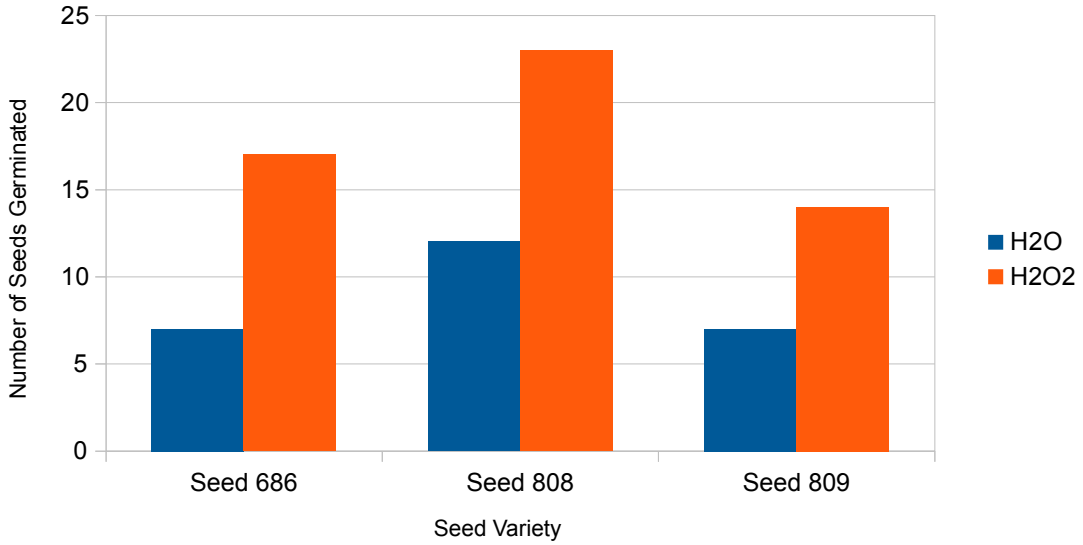


Figure 5: Effect of Hydrogen Peroxide and Water on Seed Germination

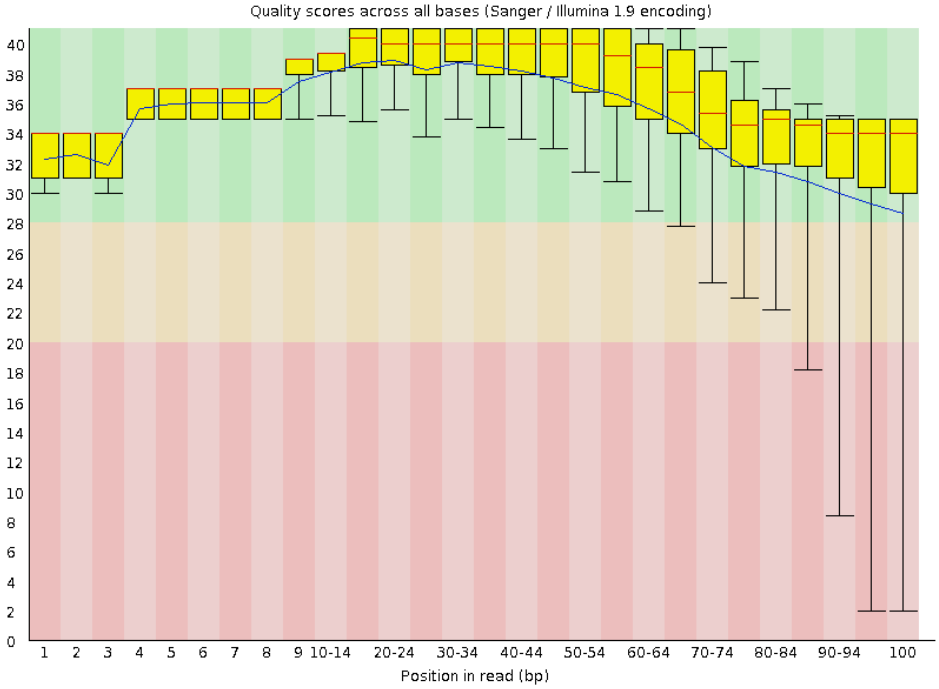


Figure 6: Quality graph across 100bp read of USH20 seed variety in H_2O solution in Phred33 format

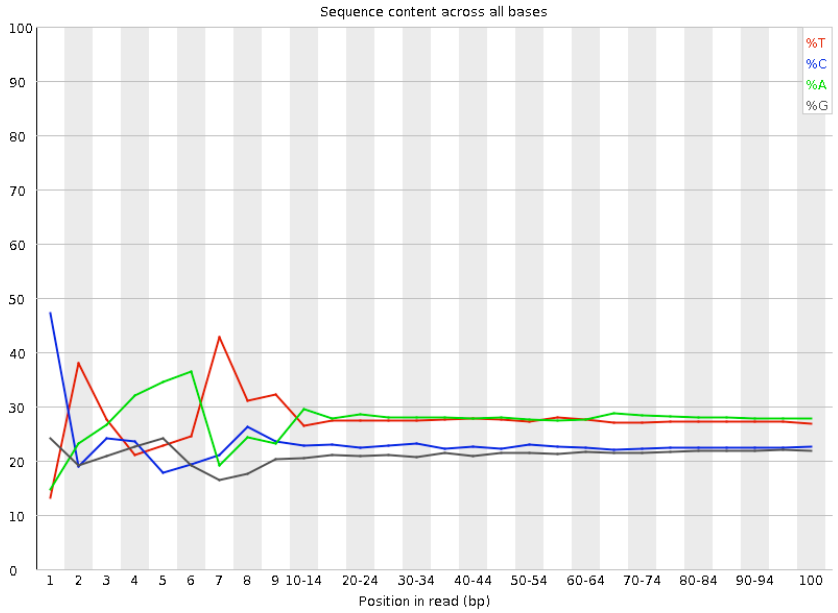


Figure 7: Graph of percentage composition of base across 100bp read; % T, % C, % A, % G

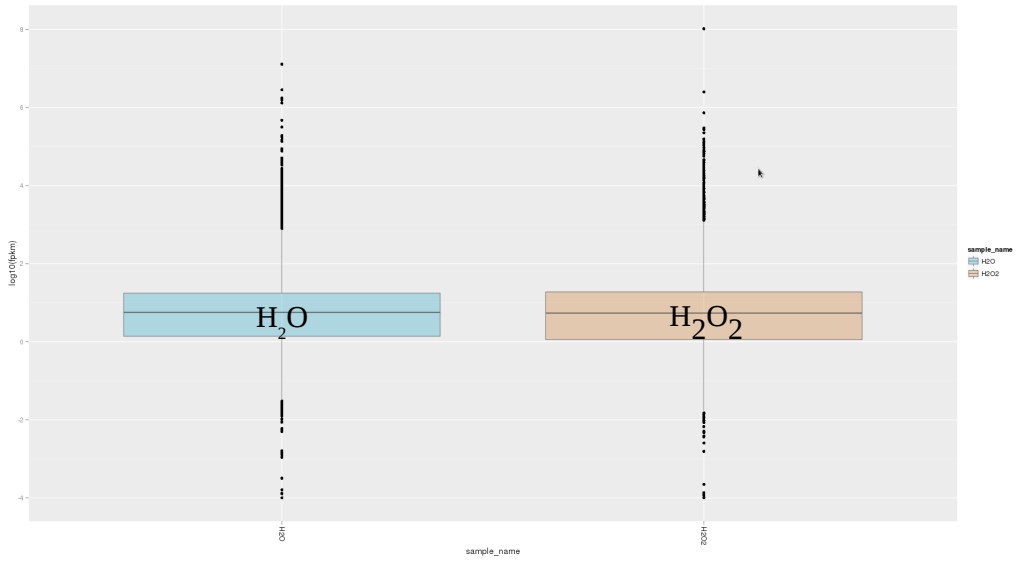
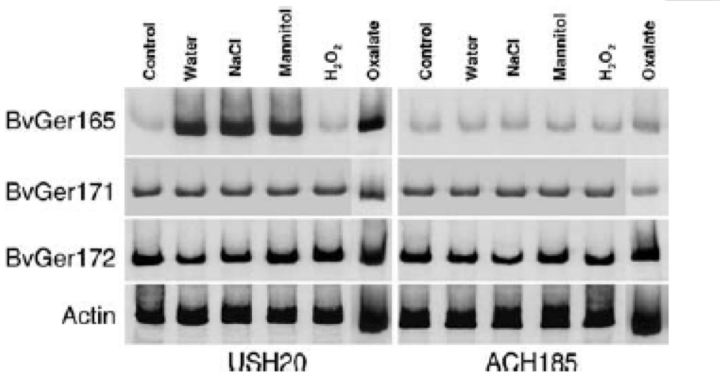


Figure 8: Box plot of Cuffdiff output for each solution type generated by CummeRbund indicates a similar distribution between both data sets



RT-PCR of USH20 (high emerger) and ACH185 (low emerger) varieties using gene-specific primers for each of the three GLP classes represented by full-length clones *BvGer*165, 171 and 172. β-actin was used as RT-PCR and a loading control

Figure 9: Expression of several genes in many solutions using the northern blot protocol

Works Cited

Bergman, Nicholas H., David Wheeler, and Medha Bhagwat. Comparative Genomics. Vol. 1-2. Totowa, NJ: Humana, 2007. Print.

De Los Reyes, BG, and JM McGrath. "Cultivar-specific Seedling Vigor and Expression of a Putative Oxalate Oxidase Germin-like Protein in Sugar Beet (*Beta vulgaris* L.)." Theoretical and Applied Genetics 107 (2003): 54-61. Print.

L. Goff and C. Trapnell (2011). cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.. R package version 1.2.0.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. "The Sequence Alignment/map (SAM) Format and SAMtools." Bioinformatics 25 (n.d.): 2078-079. Print.

McGrath, Mitchell J., Cathy Derrico, Marcos Morales, Larry Copeland, and Donald Christenson. "Germination of Sugar Beet (*Beta vulgaris* L.) Seed Submerged in Hydrogen Peroxide and Water as a Means to Discriminate Cultivar and Seedlot Vigor." Seed Science and Technology 28 (2000): 607-20. Print.

Naegele, R. P., and J. M. McGrath. "Germination and Seedling Vigor in *Beta vulgaris*." 2011 Annual Beet Sugar Development Foundation Research Report (2011): n. pag. Print.

"Sugar Beet." Fairfood International. N.p., n.d. Web. 03 Aug. 2012.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. "Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks." Nature Protocols 7 (2012): 562-78. Nature Protocols. Nature, 01 Mar. 2012. Web. 15 July 2012.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. Van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." Nature Biotechnology 28.5 (2010): 511-15. Print.