**set1Δ RNA-seq**

Nadia Atallah

February 15, 2016

**Introduction**

Set1 is an H3K4 methyltransferase in yeast. Loss of Set1 results in the loss of global histone H3K4 methylation and leads to gene expression changes at subtelomeric and euchromatic genes ([South, Harmeyer, Serratore, & Briggs, 2013](#_ENREF_16)). While previous studies have aimed to identify the genes regulated by Set1 and H3K4 methylation, these studies have failed to make a connection between loss of Set1, changes in gene expression, and a phenotype. Furthermore prior studies have failed to identify pathways that may be affected by the loss of Set1 and therefore H3K4 methylation. The Briggs lab has previously shown that conserved histone methyltransferases/epigenetic factors are necessary for thermotolerance, oxidative stress, and anaerobiosis. RNA-seq will be used to identify novel target genes and pathways that the Set1 histone methyltransferase along with set domain containing protein Set4 may regulate. Set1 was found to play a role in antisense ncRNA-mediated gene repression and set1Δ was found to affect antisense transcription ([van Dijk et al., 2011](#_ENREF_21)) and therefore an analysis to identify differential antisense transcription will also be performed.

**Methods**

**Sample Preparation, Library Preparation, and Sequencing**

Three biological replicates of each condition were sequenced, for a total of 27 samples (Table 1). If there is a difference in antisense transcription, it will not be noticeable unless we separate sense from antisense reads, therefore a strand-specific protocol was used in the library preparation. The Purdue Genomics Facility prepared libraries using the TruSeq stranded kit (Illumina, San Diego, CA) with PolyA selection. Two lanes of 2x100bp reads were sequenced using the HiSeq2500 on Rapid Run mode. Before library preparation the dscDNA quality was checked using an Agilent Bioanalyzer with the High Sensitivity DNA Chip. The *Saccharomyces cerevisiae* strain used was BY4741 (S288C background strain). Minimal media was used, in “SC” (synthetic complete) conditions. To promote oxidative stress H2O2 was utilized for some conditions.

**Table 1. Samples sequenced.**  X’s represent a biological replicate. Each condition had three biological replicates sequence.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **30°C** |  | **41°C** |
|  | Aerobic | Anaerobic | 1.75mM H2O2 | Aerobic |
| WT | X X X | X X X | X X X | X X X |
| set1Δ | X X X |  | X X X | X X X |
| set4Δ | X X X | X X X |  |  |

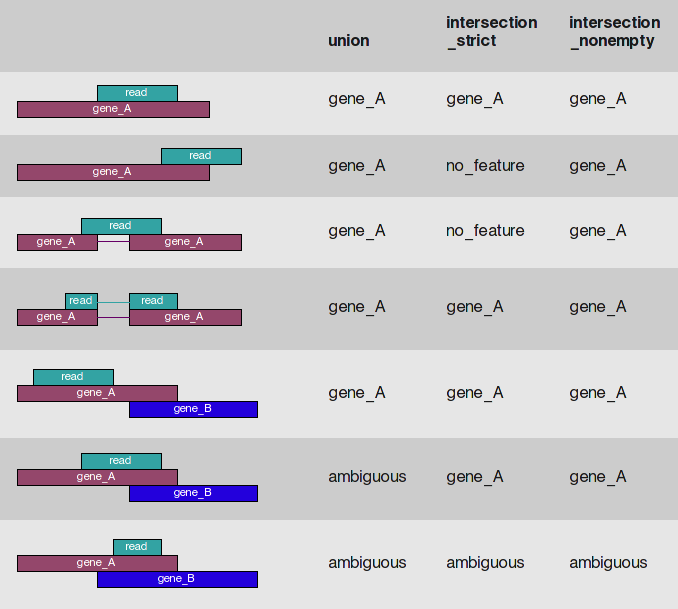
There are a number of genes which are expected to exhibit changes in expression under certain conditions used in this experiment. In *set1*Δ subtelomeric regions are expected to have gene expression changes (up-regulated) at 30°C. Based on prior qRT-PCR experiments performed by the Briggs lab, *DAN1* goes up in expression in *set1*Δ (and is not normally expressed under aerobic conditions) and *CTT1* is expected to be ~70% downregulated in *set1*Δ at 30° C. Furthermore, *DAM1*, a gene that is also methylated by Set1 and encodes for a kinetochore protein, is expected to go up in the *set1Δ* samples. Generally *DAM1* is not expressed under aerobic conditions (based on qRT-PCR)

**Adapter and Quality Trimming of Reads**

Files were downloaded and stored in the Purdue University Data Depot. Reads were quality trimmed and Illumina TruSeq adapter sequences were removed using Trimmomatic v. 0.32 ([Bolger, Lohse, & Usadel, 2014](#_ENREF_3)). Trimmomatic is a program that removes adapter sequences and trims short Illumina reads based on quality. FastQC v. 0.11.2 ([Andrews, 2010](#_ENREF_2)) was run in order to observe data quality both before and after quality trimming/adapter removal. FastX-Toolkit v. 0.0.13.2 ([Gordon, 2009](#_ENREF_5)) quality trimmer was used to further trim reads based on quality score and FASTX-Toolkit quality chart was used to make read per-base quality plots. A FastX trimscore of 30 (the minimum quality score for trimming reads was 30) and a trim length of 50 (reads shorter than 50 bases will be discarded) were used. Maximum length was set to 151 bases. All plots were checked to ensure that read quality for the reads that would be used in the remainder of the analysis were of high quality and that there were no problems with reads.

**Aligning Reads to the Reference Genome and Counting Reads**

Before any statistical analysis can be performed, reads resulting from sequencing must be aligned to a reference in order to quantify relative amounts of genes/transcripts. Tophat2 ([Kim et al., 2013](#_ENREF_8); [Trapnell, Pachter, & Salzberg, 2009](#_ENREF_18)) was used to align reads to the R64-1-1 Ensembl *Saccharomyces cerevisiae* genome release 82. Tophat2 was run with defaults except that the number of mismatches allowed was 1, max-intron-length was set to 560, min-intron-length was set to 60, and due to the strand-specificity of the library, the library-type was set to “fr-firststrand”. For the antisense analysis, the library-type was set to “fr-secondstrand”.

The htseq-count script in HTSeq v.0.6.1 ([Anders, Pyl, & Huber, 2015](#_ENREF_1)) was run to count the number of reads mapping to each gene. HTSeq used Biopython v.2.7.3 in the analysis. HTSeq was run using the GTF file with rRNA genes removed on “intersection-nonempty” mode. The HTSeq modes specify how to handle reads that overlap with more than one gene (Figure 1). The “union” mode is the default and recommended mode however this mode throws away all reads mapping to overlapping genes, even when there is evidence that suggests that the read originated from one gene over another. The HTSeq feature was set to “exon” (this specifies which feature from the GTF file is to be used). The HTSeq attribute parameter was set to “gene\_id”, which specifies that the Ensembl gene IDs are to be used as rownames in the count files. The --stranded=reverse option was set for the sense differential expression analysis.

**Figure 1.** **HTSeq modes for dealing with reads that overlap more than one feature.** Image was taken from the HTSeq manual. This image details the way in which a read is scored depending on the feature(s) the read maps to and the mode in which HTSeq is run.

The --stranded=yes option was set for the antisense differential expression analysis. Once the pipeline was completely done running, all error and output files were checked to ensure that everything ran as expected. The Purdue Bioinformatics Core’s RNA-Seq pipeline was used in trimming reads, running fastQC, running Tophat, and running HTSeq.

**Analysis using DESeq2**

The Bioconductor package DESeq2 (Love, Huber, & Anders, 2014) was used for differential expression analysis. DESeq2 incorporates estimates of dispersion and logarithmic fold changes in negative binomial generalized linear models to test for differential expression of genes. DESeq2 takes raw count data as input, in the form of a matrix with count integers. Columns are samples (libraries) and rows are gene names. These counts are then normalized and corrected to account for library size (differences in sequencing depth).

The differential expression analysis in DESeq2 uses a generalized linear model:

*Kij* ~ NB(µ*ij*, α*i*)

µ*ij* = *sj*qij

log2(q*ij*) = x*j·*β*i*

where *Kij* is the counts for gene *i* in sample *j* and the counts are modeled using a negative binomial distribution with fitted mean µ*ij* and the dispersion parameter α*i*, which is specific to every gene. Dispersions are estimated using the expected mean values from the maximum likelihood estimate of log2 fold changes. The fitted mean is the product of a size factor *sj* and the parameter qij, which is proportional to the expected true concentration of sequence fragments for sample *j*. For each column of the model matrix *X*, the *βi* coefficients give the log2 fold changes for gene *i*. The steps in the DESeq2 differential expression analysis are first to normalize the data based on the estimated size factors and the estimated dispersions, and finally to fit a negative binomial generalized linear model and to calculate Wald statistics, which are used to calculate p-values. The Benjamini-Hochberg false discovery rate correction is used to correct p-values for multiple testing.

Generally, many genes in RNA-seq data will have very low counts. These low counts tend to be highly variable, are less robust, and tend to have little or no chance of showing significant differential expression due to their high dispersion. Additionally, these genes lower the ability to detect differentially expressed genes by decreasing the power. Thus, generally in RNA-seq projects, filtering of lowly expressed genes is performed in order to increase the power and thus reduces the severity of the multiple testing adjustment made. Initially, rows with only zero counts were filtered to decrease the memory size of the DESeq data object, thus increasing the speed of DESeq2 functions. Later, independent filtering is performed by using the mean of the normalized counts as a filter statistic. A threshold is found which optimizes the adjusted p-values lower than a significance level α. Using the mean of normalized counts is independent in that it does not use any of the variables specified in the design formula. We filter out genes with low normalized counts, thus the majority of the low adjusted p-values are kept.

Count outliers are often seen in RNA-seq data. Such outliers can occur for a number of reasons, including read mapping problems, rare technical/experimental artifacts, or even rare biological events. For our purposes, we are interested in genes that show a consistent behavior within biological replicates and thus would like to ignore outliers when identifying differentially expressed genes. A diagnostic called Cook’s distance was used to test for outliers. Cook’s distance is a measure of how much a single sample is influencing the fitted coefficients for a gene - the larger the value of Cook’s distance, the more likely it is that a count is an outlier. Since the present study has three biological replicates for each condition, it was possible to calculate a Cook’s distance for each count and to flag counts which may be an outlier, based on comparison with the other biological replicates. If a count is flagged, the p-value and adjusted p-value in the results table are set to “NA”. The Cook’s distance cutoff used was the 99% quantile of the *F*(*p*,*m*-*p*) distribution, with *p* being the number of parameters including the intercept and *m* being the number of samples.

For differential expression analysis we operate on raw counts and use discrete distributions however for unsupervised data exploration it is useful to work with transformed versions of the data. The regularized logarithm, or rlog, was used in the current data analysis. The rlog function transforms the original count data to the log2 scale by fitting a model with a term for each sample and a prior distribution on the coefficients which is estimated from the data. The resulting data is of the form:

log2(*qij*) = *βi0* + *βij*

where *qij* is a parameter proportional to the expected true concentration of sequence fragments for gene *i* and sample *j*, *βi0* is an intercept (this term does not undergo shrinkage), and *βij* is the sample specific effect which is shrunk toward zero based on the dispersion-mean trend over the data. A prior on the sample specific effect terms is added. This regularized log transformation is preferable to other scaling transformations such as the variance stabilizing transformation if the size factors vary widely.

**Analysis using edgeR**

Another popular method for performing differential expression analyses on RNA-seq data is edgeR ([Robinson, McCarthy, & Smyth, 2010](#_ENREF_14)), which is highly similar to DESeq2 but uses TMM normalization ([Robinson & Oshlack, 2010](#_ENREF_15)) to scale counts into pseudocounts, on which the hypothesis tests are performed. TMM normalization (Trimmed Mean of M-values) allows for scaling factors to be calculated from the raw data and used in the ensuing statistical analysis, thus normalizing for differences in library size and also RNA composition. A trimmed mean is the average after removing the upper and lower x% of the data. The pseudocounts are used to represent the counts that would have been observed assuming the fitted model, if sequencing depths for each library had been equal. The edgeR package also uses generalized linear models to account for multifactor experimental designs ([McCarthy, Chen, & Smyth, 2012](#_ENREF_11)). In edgeR, like in DESeq2, a matrix of raw counts is the input. Then normalization factors are computed and entered into the statistical model. In the current analysis, gene-specific correction factors were entered into the functions as offsets to allow each gene to have a different dispersion.

Before hypothesis test are carried out, when fitting the negative binomial model, edgeR estimates the biological coefficient of variation (BCV), which is the coefficient of variation with which the unknown true abundance of the gene varies between replicated RNA samples. In RNA-seq experiments, the total coefficient of variation is composed of both a technical coefficient of variation and the biological coefficient of variation. The BCV is likely to be the dominant source of uncertainty for high count genes and represents the coefficient of variation that would remain if infinite sequencing depth were possible for biological replicates.

Under the negative binomial model, the reads are distributed as

*Ygi ~ NB(Mipgj, φg)*

Where *Ygijk* is the read count for gene *g*, from sample *i*, *Mi* is the total number of reads for sample *i*, *pg*is the proportion of reads mapping to gene *g* in group *j* to which sample *i* belongs, and *φg* is the gene-specific dispersion. As in DESeq2, the Benjamini-Hochberg false discovery rate correction is used to correct p-values for multiple testing.

Similarly to in the DESeq2 analysis, low count genes were removed from the edgeR analysis. Genes were kept in the analysis provided that they were present at a minimum of 1 count per million (CPM), which in this dataset corresponds to approximately 8 counts in the smallest sample (Library11\_11, which has 8,860,720 counts). Genewise dispersion estimates were used in the analysis, which allows each gene to have a different dispersion parameter, allowing for greater flexibility in the model and more power. A false discovery rate (FDR) of 5% was used as a cutoff for differential expression.

**Analysis using Cufflinks2**

A third statistical package, Cufflinks, was used in performing a differential expression analysis ([Roberts, Pimentel, Trapnell, & Pachter, 2011](#_ENREF_12); [Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011](#_ENREF_13); [Trapnell et al., 2013](#_ENREF_17); [Trapnell et al., 2012](#_ENREF_19); [Trapnell et al., 2010](#_ENREF_20)). Cufflinks2 allows differential expression analyses at both the gene level and isoform level. Furthermore, Cufflinks2 takes into account transcript length, which allows for comparison of gene expression between genes within a sample in addition to comparison of expression levels of a gene between conditions. The reason that length must be accounted for if comparing expression levels of two different genes/isoforms is that the longer a gene/transcript is, the more reads will originate from this gene/isoform and thus the higher the counts will be for this gene/isoform. DESeq2 and edgeR do not take length into consideration and thus while they provide accurate comparisons for a given gene between conditions, these packages do not allow accurate expression comparisons to be made between genes of different lengths. Initially, Cuffdiff2 transformed read alignment results into FPKM (Fragments Per Kilobase of gene model per Million fragments mapped) values and then performs tests under the assumption that the beta negative binomial model (a mixture distribution) reflects the distribution of the FPKM values. Overdispersion (the variability in gene-level counts between biological replicates) was estimated by fitting a generalized linear model through the count variance as a function of the mean. The geometric normalization method was used in the differential expression analysis with Cuffdiff, in which FPKMs are scaled via the median of the geometric means of fragment counts across all libraries, as done in DESeq (Love, Huber, & Anders, 2014).

**Seqmonk Quantification of Antisense Background**

The antisense transcription pipeline in Seqmonk v.0.32.1 was used to identify regions of antisense transcription. This pipeline takes strand-specific RNA-seq BAM files, the genome sequence, and a GTF annotation file and finds regions which show significant levels of antisense transcription. The reason this pipeline was used is that any strand-specific library will always have some mapped reads from the wrong strand, thus there will always be a low level of background of reads mapping to antisense transcripts. Seqmonk measures and quantifies the observed global antisense transcription level and then tests each gene to determine how likely it is that we would see the observed number of antisense reads in that gene by chance, generating a p-value corrected using a Benjamini and Hochberg multiple testing correction. A p-value cutoff of 0.05 was used and a list of genes passing this cutoff was compiled for each condition. The *S. cerevisiae* genome R64-1-1 was loaded into Seqmonk, along with the corresponding GTF file. All defaults were used except in the antisense pipeline, the library was set to “opposing strand”. Duplicate reads were removed in seqmonk when importing BAM files.

**Differential lncRNA Expression Analysis**

Several previous studies have identified lncRNA in yeast ([Engel et al., 2014](#_ENREF_4); [van Dijk et al., 2011](#_ENREF_21); [Xu et al., 2009](#_ENREF_22)). Siwen generously shared her scripts and method for identifying differentially expressed lncRNAs. The only modification made to the pipeline was that DESeq2 and edgeR ([Robinson et al., 2010](#_ENREF_14)) were both run to identify differentially expressed lncRNAs, instead of simply running edgeR. First, the GenomicRanges package was used to merge annotated lncRNA GTF files; ncRNAs were merged if they overlap ([Lawrence et al., 2013](#_ENREF_9)). These merged ncRNAs were annotated with genes that they overlap. It is possible for a single merged ncRNA to overlap with multiple genes. The “IntersectionNotEmpty” option was used in the summarizeOverlaps command, which is the method that was used to count reads mapping to genes using HTSeq for the sense and antisense analyses.

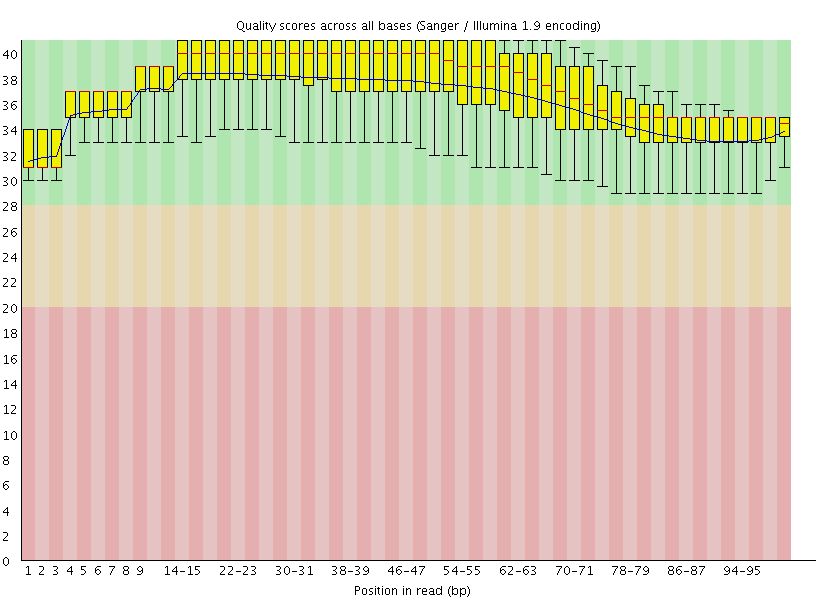
**Pathway and GO enrichment Analyses**

DAVID (the Database for Annotation, Visualization, and Integrated Discovery) was used in the annotation of genes and in performing the GO (gene ontology) enrichment analyses ([Huang da, Sherman, & Lempicki, 2009a](#_ENREF_6), [2009b](#_ENREF_7)). The analyses were performed for differentially expressed genes found using all three statistical packages as well as antisense transcripts found to be differentially expressed using all three statistical packages.

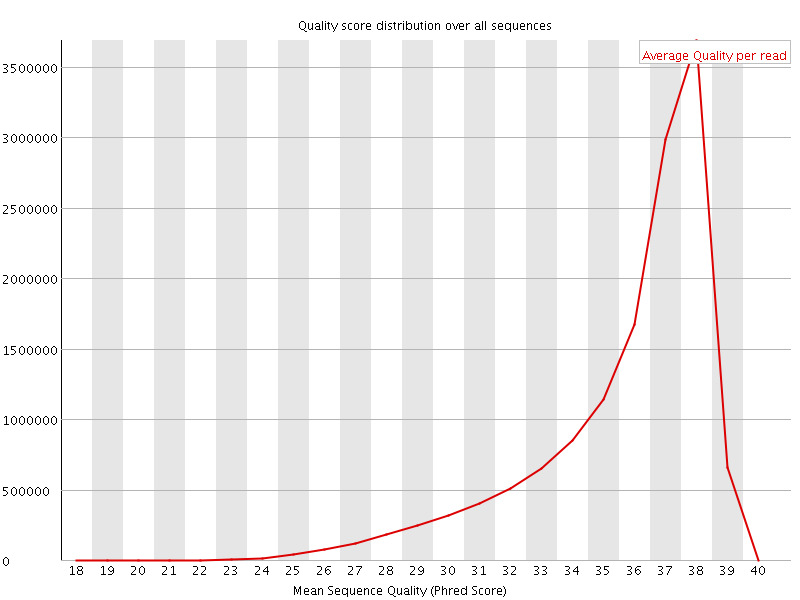
**Results**

**Quality Control Filtering and Data Visualization of FASTQ Files Shows that Data is of Hiqh Quality**

FastQC ([Andrews, 2010](#_ENREF_2)) is a program that provides simple visual quality checks for data. FastQC was originally developed for DNA sequencing data and thus not all the plots are useful for RNA-Seq analysis. The FastQC output that is the most useful for RNA-Seq data are the per base sequence quality graphs (Figure 2), the per sequence quality score graphs (Figure 3), and a list given for each FASTQ file of overrepresented sequences. All FastQC output was reviewed and the data looked to be of high quality. The total number of reads sequenced for all 27 samples was 731,850,370 reads (an average of 27,105,569 reads per sample). The total number of reads remaining post-trimming with Trimmomatic is 686,377,538 and the average number is 25,421,390. Overall there were 27 samples sequenced, all of which were sequenced successfully. After the data was trimmed using FastX-Toolkit, the number of total reads remaining is 671,257,282 and the average number of reads remaining per sample is 24,861,381. Due to the high quality of the sequencing data, 91.72% of reads remained after trimming. No samples had higher than 10% of the reads match to rRNA. The sample with the highest rRNA content was sample 10\_3 (WT in H2O2 Replicate 1) with only 5.6% of reads mapping to rRNA, which is still low. Overall the data quality appears high.



**Figure 2. Per base sequence quality for *set1Δ* at 41°C replicate 1 trimmed left reads.** The y-axis shows quality scores and the x-axis shows the position in the read. The red line is the median and the blue line is the mean. The yellow box represents the inter-quartile (25%-75%) range, and the upper and lower whiskers represent the 10% and 90% points. The dip in sequence quality at the end of the read is normal and is nearly always observed in Illumina sequencing data. Due in part to trimming, all positions in the reads have high quality. This graph is very typical for what was observed across all read files.



**Figure 3.** **Per sequence quality score for *set1Δ* at 41°C replicate 1 trimmed left reads.** If a subset of sequences in the FASTQ file have universally low quality scores it will show up in this plot. The Y axis shows the number of sequences and the x axis shows the mean quality score. This plot shows that the reads post-trimming have very high quality. This plot looks very similar to all the other per sequence quality score plots seen across all the trimmed FASTQ files.

**96% of Reads Map to 7,110 Features**

Tophat2 ([Kim et al., 2013](#_ENREF_8)) was successfully run on all samples and overall 645,277,458 reads were mapped out of the 671,257,282 input reads (~96%). Subsequently, HTSeq ([Anders et al., 2015](#_ENREF_1)) was run to count reads aligning to 7,110 features (coding genes, pseudogenes, and non-coding genes) in the GTF annotation file.

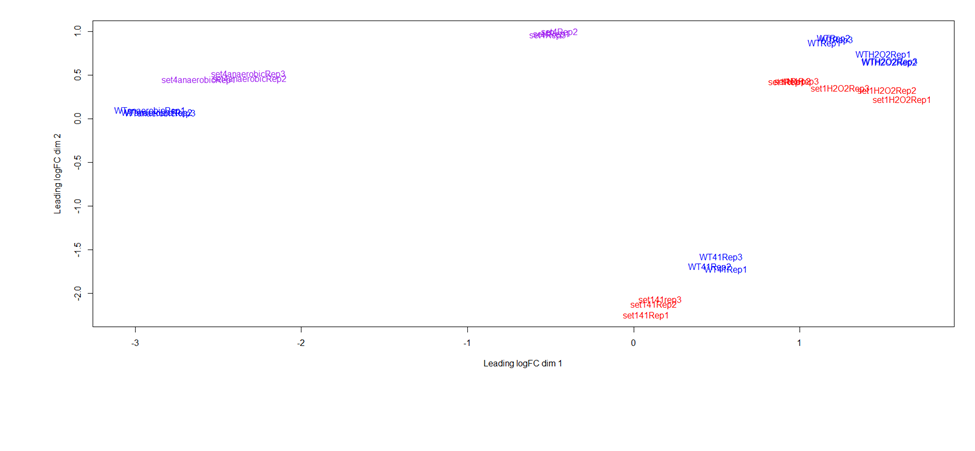
Ribosomal RNA (rRNA) can often cause issues in RNA-Seq data. If no RNA selection is done prior to library preparation, up to ~95% of the reads can be expected to map to rRNA. For the current experiment, PolyA selection was performed, and therefore polyadenylated RNA were selected. This practice excludes the majority of rRNA. Unfortunatley polyA selection is not 100% efficient at getting rid of rRNA and so there is always some rRNA contamination in sequencing reads. The polyA selection seems to have worked very well, and none of the 27 libraries used in the analysis had greater than 10% of rRNA contamination. Since reads were being aligned to a reference genome, the presence of rRNA reads is not a huge concern. However it is plausible that the uneven distribution of rRNA reads amongst the libraries could throw off the statistical analysis. Therefore steps should be taken to remove rRNA prior to statistical analysis. There are two simple ways to remove these rRNA genes and associated counts prior to performing the statistics: one is to remove the rRNA genes based on Ensembl gene ID in R and the second method is to remove the rRNA genes from the GTF file prior to running HTSeq ([Anders et al., 2015](#_ENREF_1)). These two methods produce identical results. The count matrix generated from the GTF file without rRNA will not have rRNA genes in it and will have the rRNA reads counted in the “Map to No Feature” Category. HTSeq ([Anders et al., 2015](#_ENREF_1)) was run in intersection-nonempty mode (see Figure 1 for details on HTSeq modes).

**Genes with Zero Counts were filtered**

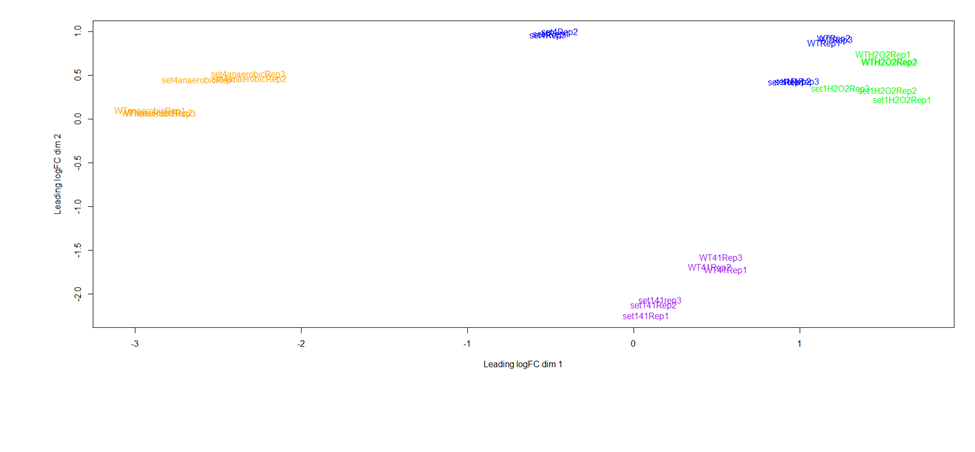
The original data were comprised of 7,110 features. Features with zero counts across all libraries were filtered, thus increasing the statistical power and the ability to detect truly differentially expressed genes. After removing features with zero counts, 6,930 features remained, or ~97% of the original features.

**Exploratory plots were generated**

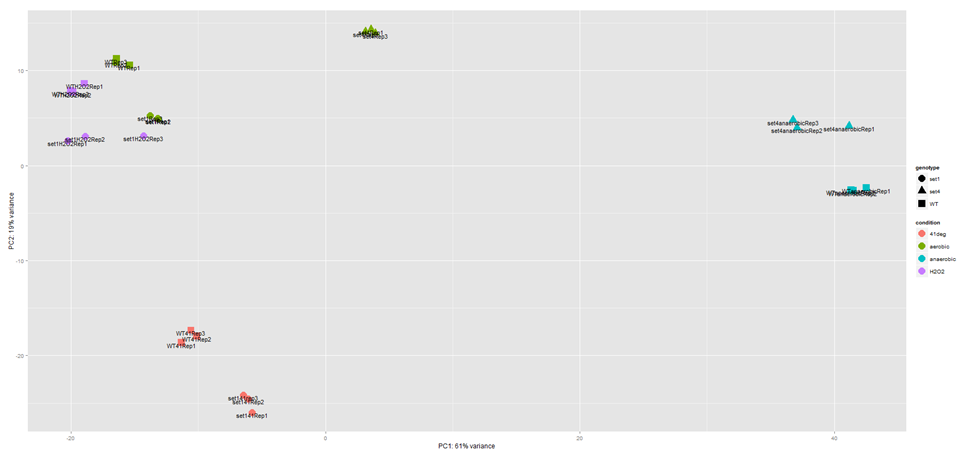
Normalization was performed using DESeq2 and normalized counts are given in the table BriggsDESeq2NormalizedCounts.csv. Once data has been normalized, it is useful to generate exploratory plots to get an idea of how samples are clustering, general attributes of the data, and whether any libraries are suspect and should be removed prior to performing the differential expression analysis. A multidimensional scaling (MDS) plot on TMM normalized data was generated (Figure 3 and Figure 4). MDS plots are a useful way to visually assess the similarities and dissimilarities between samples. The distance between each pair of samples is the Euclidean distance for genes with the highest log fold change between those samples. Therefore, samples that are similar to each other will group together. These plots suggest a clear degree of biological difference between conditions as well as based on genotype. In addition to the MDS plots, a principal components analysis (PCA) was performed (Figure 3); PCA is another way to reduce the dimensionality of data and observe clustering of samples. In the MDS plots, the PCA plot, and a heatmap of the 90 most highly expressed genes (Figure 7), samples seem to have a high degree of similarity within biological replicates.



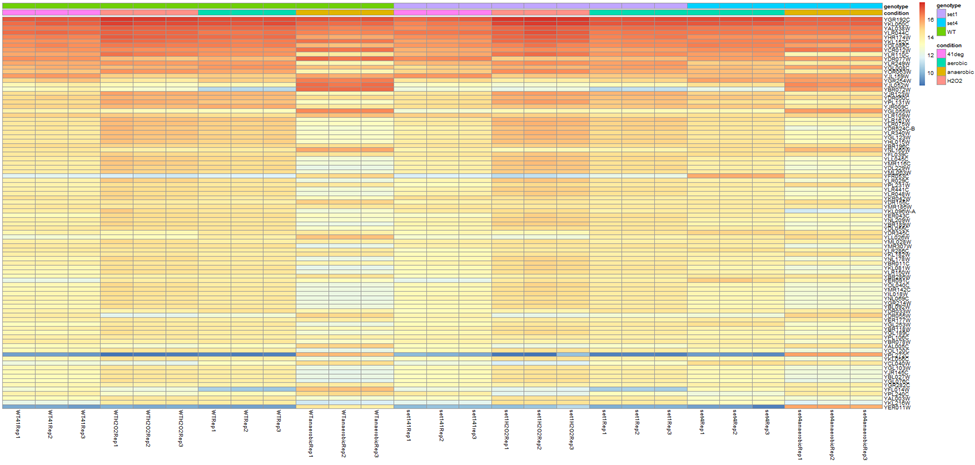
**Figure 4. MDS plot colored according to genotype.** In this plot, WT is shown in blue, *set1Δ* is shown in red, and *set4Δ* is shown in purple.



**Figure 5. MDS plot colored by condition.** Growth at 41°C is shown in purple, growth under oxidative stress (H2O2) is shown in green, normal growth conditions are shown in blue, and anaerobic conditions are shown in orange.

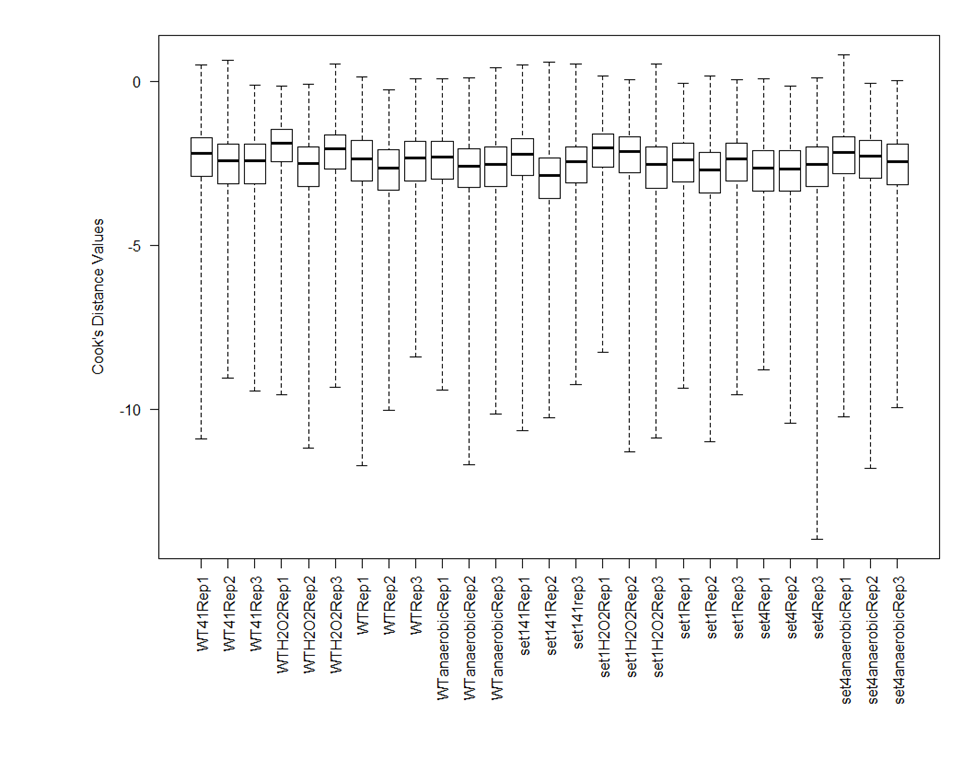


**Figure 6. PCA Plot.** This plot shows the samples separated in a 2D plane spanned by the first two principal components. The first principal component accounts for ~61% of the overall variability, whereas the second principal component accounts for ~19% of the overall variability. Genotype is represented by shape, with *set1Δ* as circles, *set4Δ* as triangles, and WT as squares. Condition is represented by various colors, with salmon representing 41°C conditions, green representing standard aerobic conditions, teal representing anaerobic conditions, and lavender representing oxidative stress conditions (the presence of H2O2). Samples appear to be clustering based on both condition and genotype.

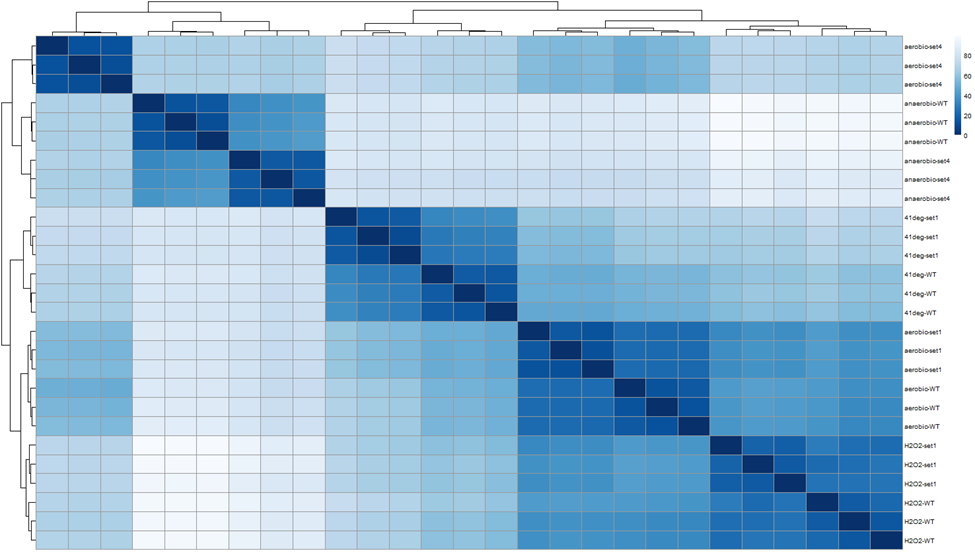


**Figure 7. Heatmap showing the expression data of the 90 most highly expressed genes.** The data shows the normalized counts after having a regularized log transformation performed using DESeq2.

Cook’s distance is a measure of how much a single sample is influencing the fitted coefficients for a gene - and is a useful way to detect outliers in experiments with at least three biological replicates per condition. Genes in the results tables which have p-values and adjusted p-values listed as “NA” were flagged as outliers based on Cook’s distance. In Figure 8, boxplots show the distribution of Cook’s distances for each library. Figure 9 shows the samples clustered based on Euclidean distances of regularized log transformed data. In Figure 9, it is clear that samples are clustering based first on condition and second based on genotype. All biological replicates are clustering together as expected. Based on Figures 8 and 9, there are no libraries which seem to be outliers or suspect.

****

**Figure 8. Boxplots of Cook’s Distances.** Here, boxplots are used to determine whether one sample has much higher Cook’s distances than other samples - a useful methods for detecting outlier samples. In this data, all samples have a similar range of Cook’s distances.

****

**Figure 9. Heatmap of sample-to-sample distances.** A heatmap showing the Euclidian distances between samples, made with the transformed data after a regularized log transformation was performed.

**A differential expression analysis was performed to find differentially expressed genes using DESeq2, edgeR, and Cufflinks**

DESeq2 ([Love, Huber, & Anders, 2014](#_ENREF_10)), edgeR ([Robinson et al., 2010](#_ENREF_14); [Robinson & Oshlack, 2010](#_ENREF_15)), and Cufflinks (Roberts, Pimentel, Trapnell, & Pachter, 2011; Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Trapnell et al., 2013; Trapnell et al., 2012; Trapnell et al., 2010) were used to perform a differential expression analysis. Controlling the FDR at 1%, we found a number of differentially expressed genes when we compared various conditions and genotypes. These differential expression results are summarized in Tables 2,3 and 4, as well as in the MA plots shown in Figure 10. Figure 11 shows the venn diagrams summarizing the overlap between the three packages. Overall the three packages show a high degree of agreement between differentially expressed gene lists.

**Table 2. DESeq2 differential expression results.** The number of genes found to be differentially expressed using DESeq2 at a 1% false discovery rate are shown, along with the number of genes differentially expressed when an additional 2 fold change cutoff is applied. The number of genes up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

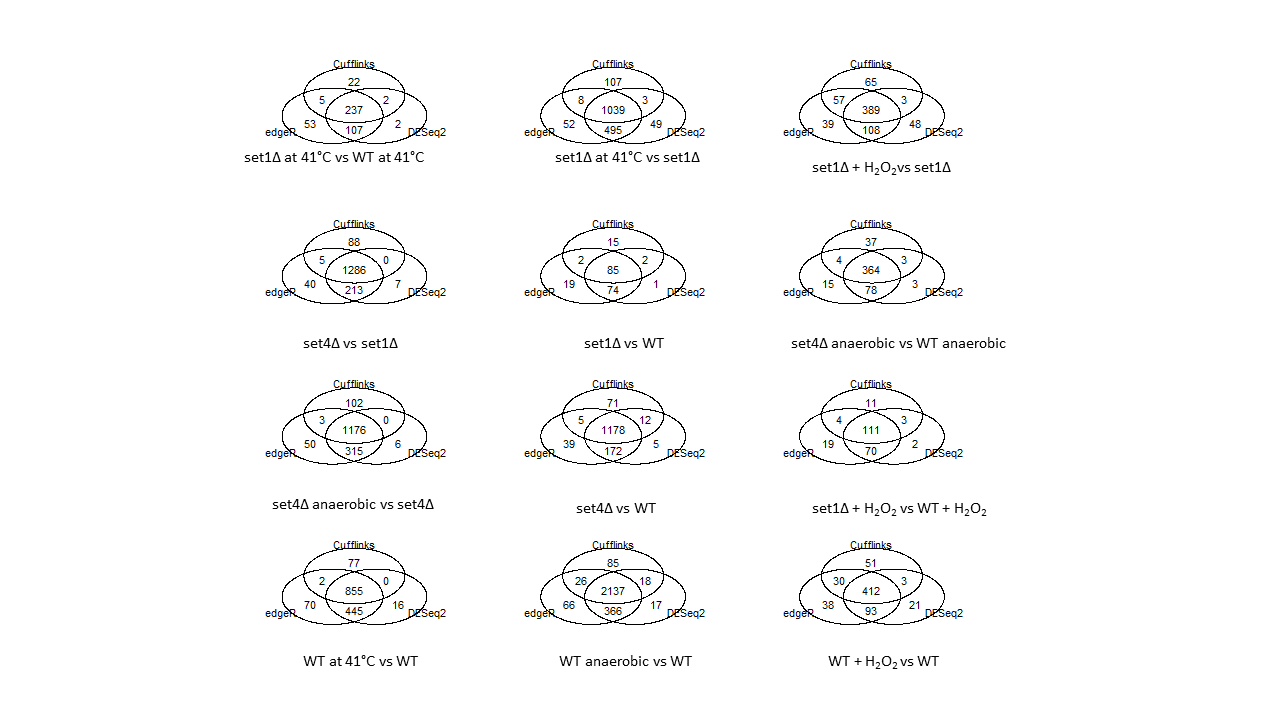
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 4727 | 2538 | 1282 | 1256 |
| *set1*Δ vs WT | 1146 | 162 | 149 | 13 |
| *set4*Δ vs *set1*Δ | 4045 | 1506 | 655 | 851 |
| *set4*Δ vs WT | 3891 | 1367 | 653 | 714 |
| *set1*Δ at 41°C vs set1 | 4382 | 1586 | 1084 | 502 |
| *set4*Δ anaerobic vs *set4*Δ | 4123 | 1497 | 848 | 649 |
| *set4*Δ anaerobic vs WT anaerobic | 2590 | 448 | 135 | 313 |
| WT + H2O2 vs WT | 3303 | 529 | 169 | 360 |
| *set1*Δ + H2O2 vs *set1*Δ | 3405 | 548 | 147 | 401 |
| *set1*Δ + H2O2 vs WT + H2O2 | 1466 | 186 | 119 | 67 |
| *set1*Δ at 41°C vs WT at 41°C | 2222 | 348 | 318 | 30 |
| WT at 41°C vs WT | 4108 | 1316 | 855 | 461 |

**Table 3. edgeR differential expression results.** The number of genes found to be differentially expressed using edgeR at a 1% false discovery rate are shown, along with the number of genes differentially expressed when an additional 2 fold change cutoff is applied. The number of genes up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

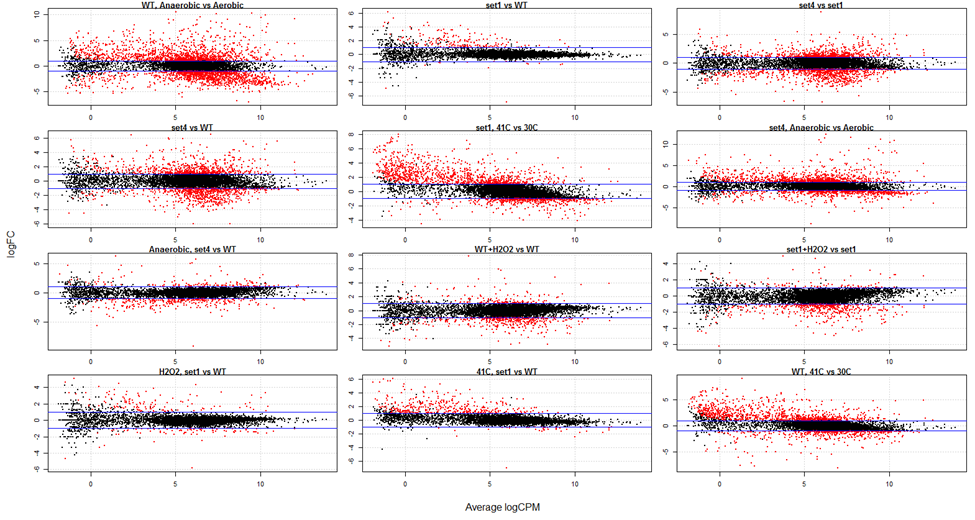
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 4779 | 2595 | 1276 | 1319 |
| *set1*Δ vs WT | 1154 | 180 | 163 | 17 |
| *set4*Δ vs *set1*Δ | 4101 | 1544 | 673 | 871 |
| *set4*Δ vs WT | 3940 | 1394 | 649 | 745 |
| *set1*Δ at 41°C vs set1 | 4417 | 1594 | 1077 | 517 |
| *set4*Δ anaerobic vs *set4*Δ | 4195 | 1544 | 870 | 674 |
| *set4*Δ anaerobic vs WT anaerobic | 2647 | 461 | 149 | 312 |
| WT + H2O2 vs WT | 3295 | 573 | 151 | 422 |
| *set1*Δ + H2O2 vs *set1*Δ | 3407 | 593 | 100 | 493 |
| *set1*Δ + H2O2 vs WT + H2O2 | 1471 | 204 | 124 | 80 |
| *set1*Δ at 41°C vs WT at 41°C | 2378 | 402 | 366 | 36 |
| WT at 41°C vs WT | 4201 | 1372 | 896 | 476 |

**Table 4. Cufflinks differential expression results.** The number of genes found to be differentially expressed using Cufflinks at a 1% false discovery rate are shown, along with the number of genes differentially expressed when an additional 2 fold change cutoff is applied. The number of genes up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 4128 | 2266 | 1071 | 1195 |
| *set1*Δ vs WT | 637 | 104 | 88 | 12 |
| *set4*Δ vs *set1*Δ | 3665 | 1379 | 608 | 771 |
| *set4*Δ vs WT | 3463 | 1266 | 620 | 646 |
| *set1*Δ at 41°C vs set1 | 3585 | 1157 | 634 | 523 |
| *set4*Δ anaerobic vs *set4*Δ | 3494 | 1281 | 658 | 623 |
| *set4*Δ anaerobic vs WT anaerobic | 1982 | 408 | 146 | 262 |
| WT + H2O2 vs WT | 2327 | 496 | 85 | 381 |
| *set1*Δ + H2O2 vs *set1*Δ | 2525 | 514 | 65 | 449 |
| *set1*Δ + H2O2 vs WT + H2O2 | 519 | 129 | 65 | 64 |
| *set1*Δ at 41°C vs WT at 41°C | 1490 | 266 | 232 | 34 |
| WT at 41°C vs WT | 3181 | 934 | 561 | 373 |



**Figure 11. Venn diagrams of overlap between differential expression results amongst the three statistical packages.**

****

**Figure 10. MA plots for each of the comparisons made during the differential expression analysis.** These plots, made using DESeq2, show the log2 fold changes on the y axis and the mean of normalized counts on the x-axis. Genes differentially expressed at a 1%FDR are colored in red.

**A differential expression analysis was performed to find differentially expressed antisense transcripts using DESeq2, edgeR, and Cufflinks**

DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010; Robinson & Oshlack, 2010), and Cufflinks (Roberts, Pimentel, Trapnell, & Pachter, 2011; Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Trapnell et al., 2013; Trapnell et al., 2012; Trapnell et al., 2010) were also used to perform a differential expression analysis on antisense transcripts. A total of 6,839 features were found with antisense transcription. Controlling the FDR at 1%, we found a number of differentially expressed antisense transcripts when we compared various conditions and genotypes. These differential expression results are summarized in Tables 5,6 and 7. Because there is a low level of reads mapping to antisense transcripts that always occurs even in strand-specific RNA-Seq protocols, the program Seqmonk was run as well. Seqmonk allows for the quantification of background antisense transcripts present (thus, transcripts that are not necessarily antisense, but are quite possibly observed due to the fact that even strand-specific library preparation protocols are not perfectly strand-specific). Table 8 gives lists the number of regions found to likely have *bona fide* antisense transcripts present, at a level larger than would be expected based simply on background noise, at α=0.05. Differentially expressed antisense results are given for each of the three statistical methods, as well as for the intersection of all three statistical methods, and the intersection of all three statistical methods for only those transcripts found to have significant antisense expression by Seqmonk. To obtain a list of the intersection between all three statistical packages and Seqmonk, a list of all observed antisense transcripts found by Seqmonk was obtained by merging the Seqmonk results for all six samples used in a given comparison.

The genes *DAN1*, *DAM1*, and *CTT1* are expected to change in some of the conditions assayed in this experiment, thus plots showing normalized counts of these genes in each condition were made, both for sense and antisense analyses (Figures 13-18). As expected, *DAN1* is upregulated in WT anaerobic conditions compared to in aerobic conditions. This upregulation of *DAN1* under anaerobic conditions is both observable in normalized counts (Figure 13) and is found to be statistically significant by all three statistical packages used. *DAN1* was also expected to be up upregulated in *set1Δ* mutants, however the gene was not found to be statistically significant. However upon examination of the scatter plots (Figures 13 and 14), it appears that while the sense transcript is not differentially expressed in *set1Δ* mutants compared to WT, that the antisense transcript is differentially expressed. Indeed, both edgeR and DESeq2 found the *DAN1* antisense transcript to be differentially expressed, although Cufflinks did not; for this reason, and due to the low level of overlap between Cufflinks with edgeR and DESeq2 (Figure 12), we may want to disregard the Cufflinks output for the antisense transcripts. The *DAM1* counts were also plotted for both sense (Figure 15) and antisense (Figure 16) transcripts. Surprisingly, not only is *DAM1* not found to be differentially expressed, it is also observed in aerobic conditions. *CTT1* was also plotted (Figures 17 and 18) and appears to be down a bit in *set1Δ* at 30°, however it was not found to be statistically significant by any package at a 1% FDR.

**Table 5. DESeq2 differentially expressed antisense transcripts.** The number of genes found to be differentially expressed by DESeq2 at a 1% false discovery rate are shown, along with the number of antisense differentially expressed when an additional 2 fold change cutoff is applied. The number of antisense transcripts up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 2295 | 1964 | 1121 | 843 |
| *set1*Δ vs WT | 234 | 153 | 123 | 30 |
| *set4*Δ vs *set1*Δ | 1223 | 902 | 420 | 482 |
| *set4*Δ vs WT | 1031 | 702 | 365 | 337 |
| *set1*Δ at 41°C vs set1 | 2362 | 1975 | 1180 | 795 |
| *set4*Δ anaerobic vs *set4*Δ | 1629 | 1257 | 752 | 505 |
| *set4*Δ anaerobic vs WT anaerobic | 510 | 263 | 97 | 166 |
| WT + H2O2 vs WT | 569 | 270 | 116 | 154 |
| *set1*Δ + H2O2 vs *set1*Δ | 520 | 256 | 126 | 130 |
| *set1*Δ + H2O2 vs WT + H2O2 | 205 | 140 | 102 | 38 |
| *set1*Δ at 41°C vs WT at 41°C | 891 | 512 | 417 | 95 |
| WT at 41°C vs WT | 1879 | 1496 | 906 | 590 |

**Table 6. edgeR differentially expressed antisense transcripts.** The number of genes found to be differentially expressed by edgeR at a 1% false discovery rate are shown, along with the number of antisense transcripts differentially expressed when an additional 2 fold change cutoff is applied. The number of antisense transcripts up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

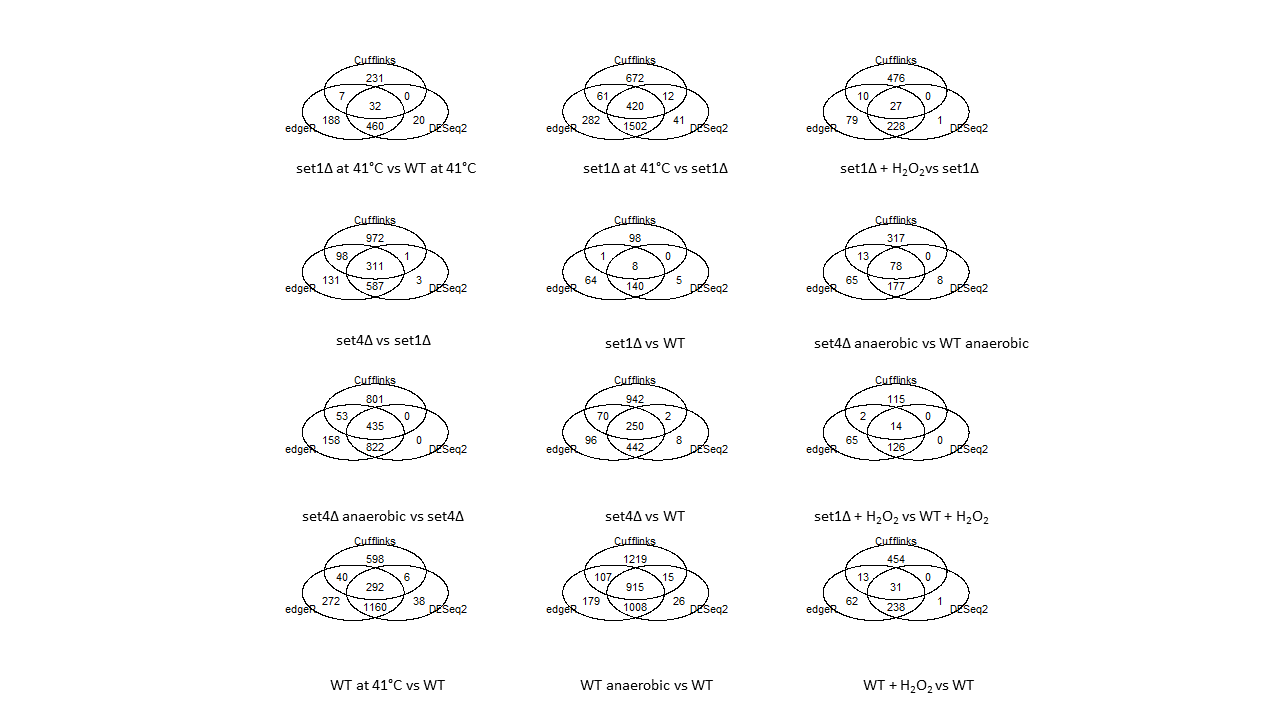
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 2578 | 2209 | 1160 | 1049 |
| *set1*Δ vs WT | 328 | 213 | 157 | 56 |
| *set4*Δ vs *set1*Δ | 1509 | 1127 | 535 | 592 |
| *set4*Δ vs WT | 1228 | 858 | 410 | 448 |
| *set1*Δ at 41°C vs set1 | 2744 | 2265 | 1262 | 1003 |
| *set4*Δ anaerobic vs *set4*Δ | 1902 | 1468 | 826 | 642 |
| *set4*Δ anaerobic vs WT anaerobic | 634 | 333 | 143 | 190 |
| WT + H2O2 vs WT | 649 | 344 | 147 | 197 |
| *set1*Δ + H2O2 vs *set1*Δ | 618 | 344 | 180 | 164 |
| *set1*Δ + H2O2 vs WT + H2O2 | 285 | 207 | 141 | 66 |
| *set1*Δ at 41°C vs WT at 41°C | 1317 | 687 | 535 | 152 |
| WT at 41°C vs WT | 2214 | 1764 | 963 | 801 |

**Table 7. Cufflinks differentially expressed antisense transcripts.** The number of genes found to be differentially expressed by Cufflinks at a 1% false discovery rate are shown, along with the number of antisense transcipts differentially expressed when an additional 2 fold change cutoff is applied. The number of antisense transcripts up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

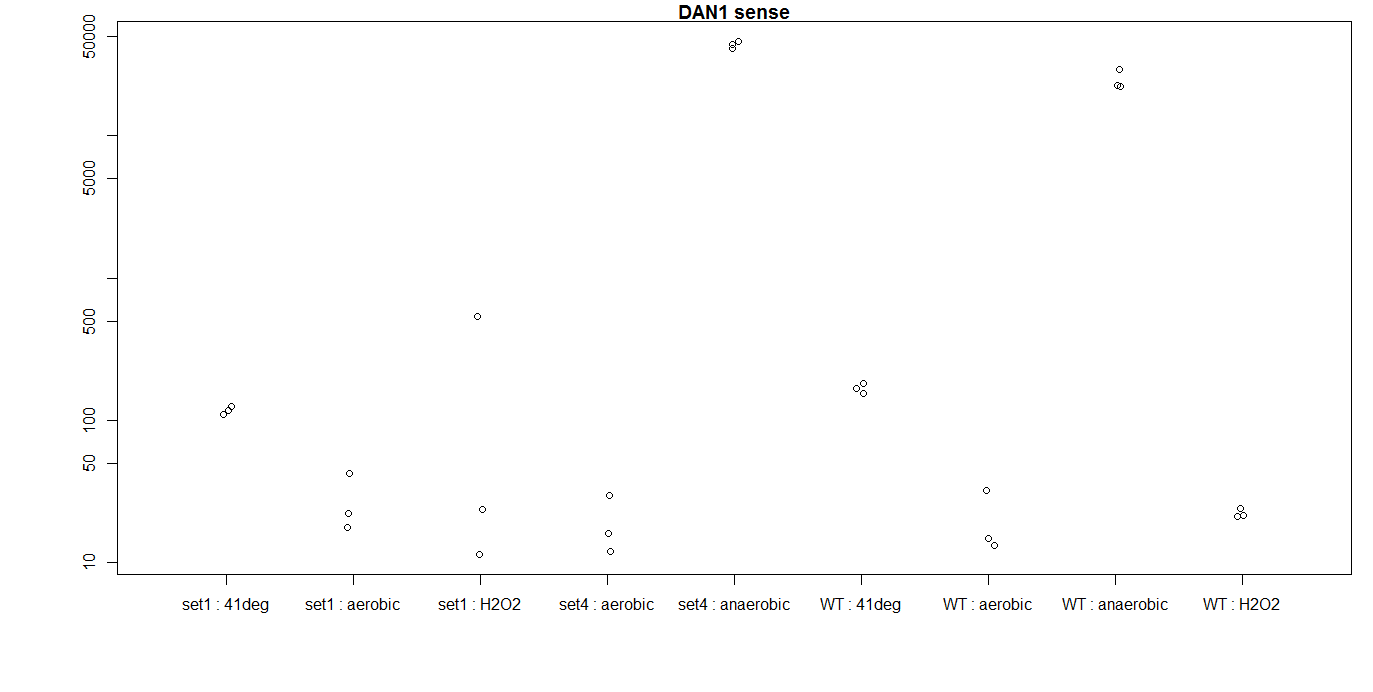
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 4119 | 2256 | 1058 | 1198 |
| *set1*Δ vs WT | 634 | 108 | 95 | 13 |
| *set4*Δ vs *set1*Δ | 3677 | 1382 | 608 | 774 |
| *set4*Δ vs WT | 3473 | 1264 | 617 | 647 |
| *set1*Δ at 41°C vs set1 | 3591 | 1165 | 639 | 526 |
| *set4*Δ anaerobic vs *set4*Δ | 3518 | 1289 | 741 | 548 |
| *set4*Δ anaerobic vs WT anaerobic | 1977 | 408 | 146 | 262 |
| WT + H2O2 vs WT | 2331 | 498 | 119 | 379 |
| *set1*Δ + H2O2 vs *set1*Δ | 2544 | 513 | 67 | 446 |
| *set1*Δ + H2O2 vs WT + H2O2 | 533 | 131 | 68 | 63 |
| *set1*Δ at 41°C vs WT at 41°C | 1502 | 270 | 232 | 38 |
| WT at 41°C vs WT | 3187 | 936 | 563 | 373 |

**Table 8. Number of regions with antisense transcription, according to Seqmonk.**

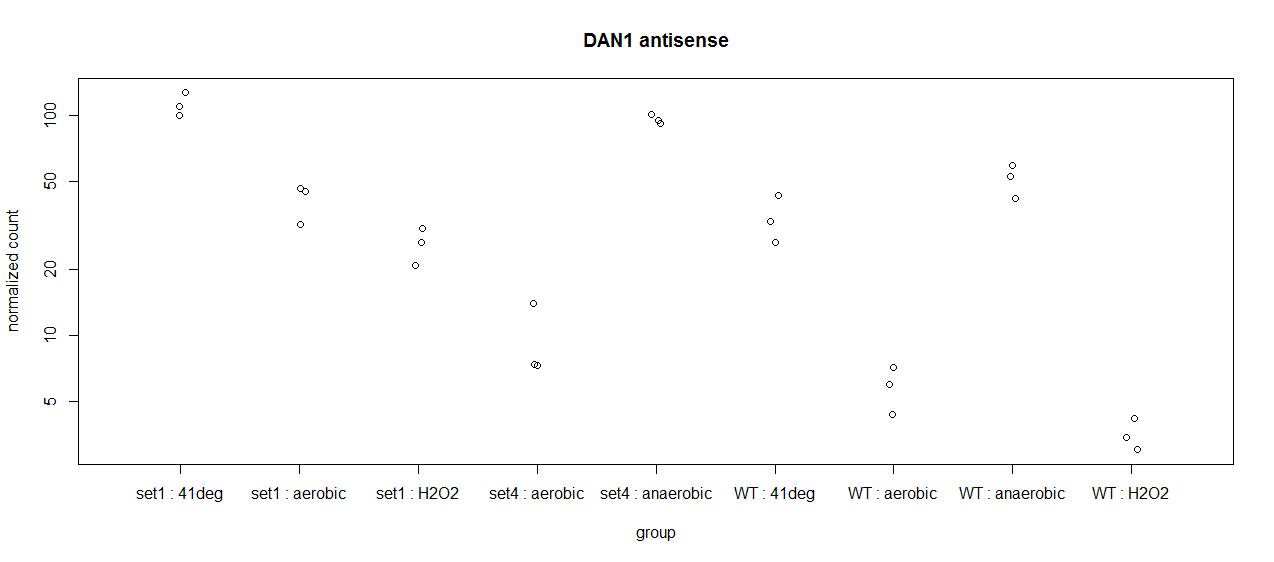
|  |  |
| --- | --- |
| **Sample** | **Regions with antisense transcription at FDR<0.05** |
| set1Δ+H2O2 Replicate 1 | 3042 |
| set1Δ+H2O2 Replicate 2 | 2874 |
| set1Δ+H2O2 Replicate 3 | 2884 |
| set1Δ at 41°C Replicate 1 | 4493 |
| set1Δ at 41°C Replicate 2 | 4291 |
| set1Δ at 41°C Replicate 3 | 4178 |
| set1Δ Replicate 1 | 2965 |
| set1Δ Replicate 2 | 2929 |
| set1Δ Replicate 3 | 2876 |
| set4Δ Replicate 1 | 3092 |
| set4Δ Replicate 2 | 3014 |
| set4Δ Replicate 3 | 2996 |
| WT Replicate 1 | 2906 |
| WT Replicate 2 | 2894 |
| WT Replicate 3 | 3004 |
| set4Δ anaerobic Replicate 1 | 3422 |
| set4Δ anaerobic Replicate 2 | 3206 |
| set4Δ anaerobic Replicate 3 | 3287 |
| WT at 41°C Replicate 1 | 3927 |
| WT at 41°C Replicate 2 | 3710 |
| WT at 41°C Replicate 3 | 3684 |
| WT anaerobic Replicate 1 | 3345 |
| WT anaerobic Replicate 2 | 3302 |
| WT anaerobic Replicate 3 | 3279 |
| WT+H2O2 Replicate 1 | 2793 |
| WT+H2O2 Replicate 2 | 2945 |
| WT+H2O2 Replicate 3 | 2841 |



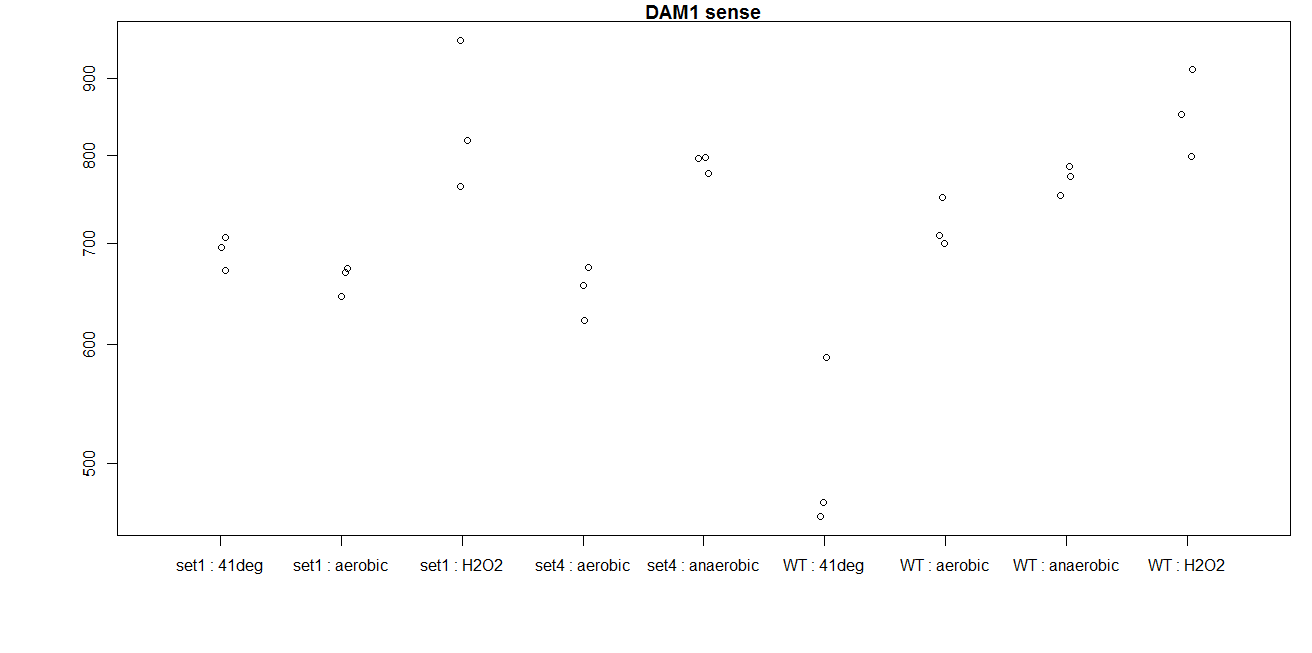
**Figure 12. Venn diagrams of overlap between antisense differential expression results amongst the three statistical packages.**



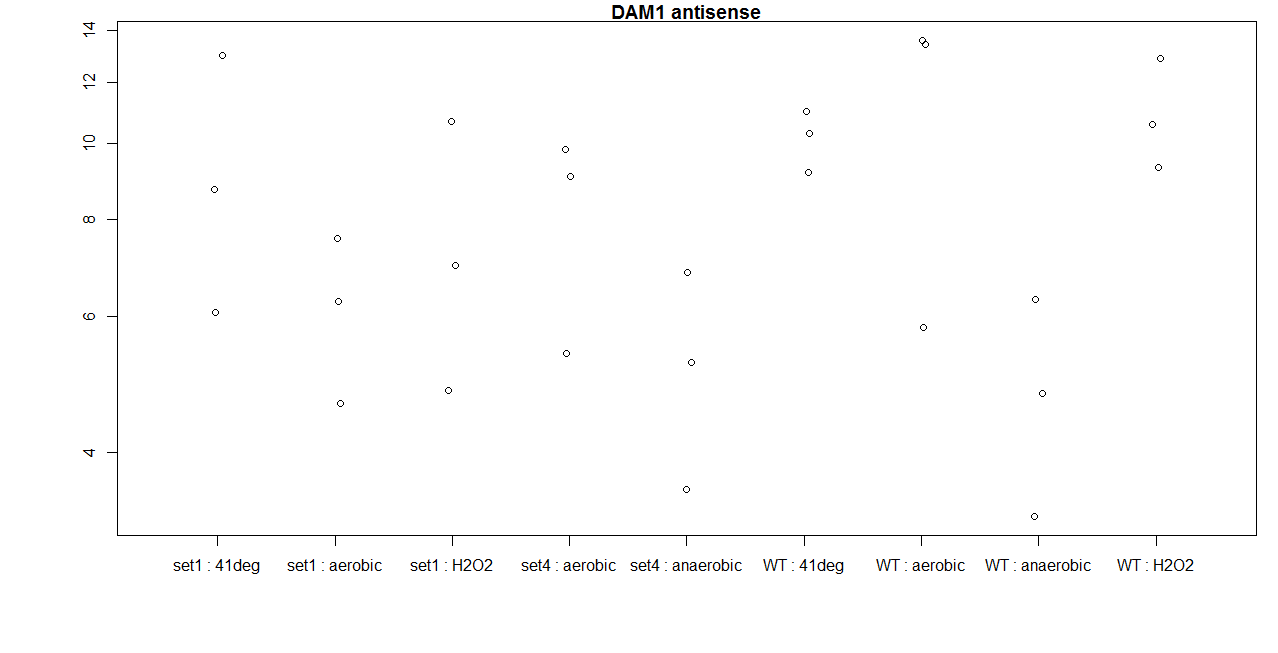
**Figure 13. Normalized counts for the gene encoding for sense DAN1.** The x-axis shows the groups and the y-axis shows the normalized counts.



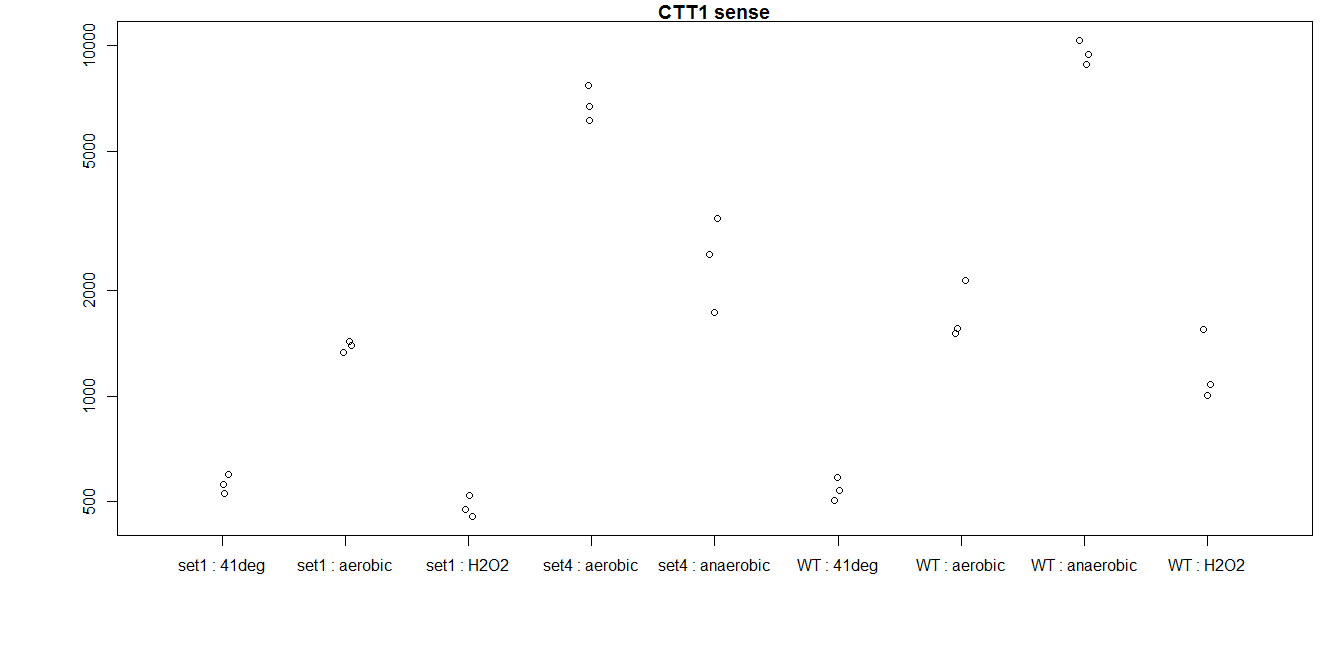
**Figure 14. Normalized counts for the gene encoding for antisense DAN1.** The x-axis shows the groups and the y-axis shows the normalized counts.



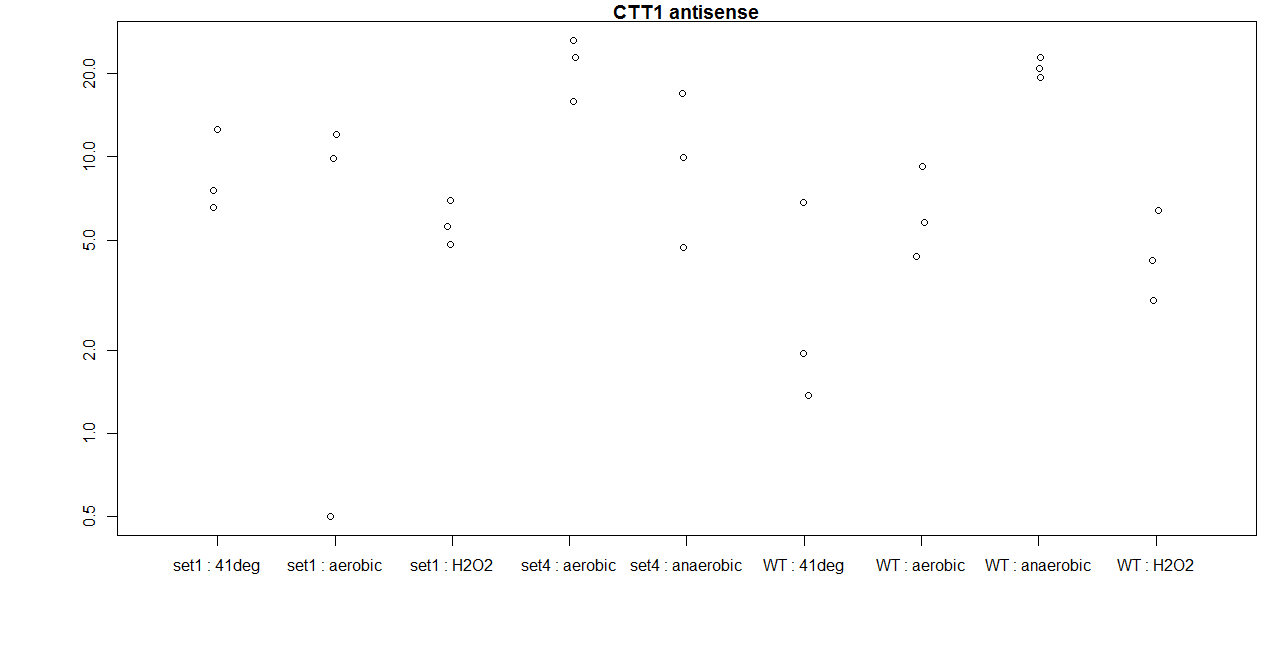
**Figure 15. Normalized counts for the gene encoding for sense DAM1.** The x-axis shows the groups and the y-axis shows the normalized counts.



**Figure 16. Normalized counts for the gene encoding for antisense DAM1.** The x-axis shows the groups and the y-axis shows the normalized counts.



**Figure 17. Normalized counts for the gene encoding for sense CTT1.** The x-axis shows the groups and the y-axis shows the normalized counts.



**Figure 18. Normalized counts for the gene encoding for antisense CTT1.** The x-axis shows the groups and the y-axis shows the normalized counts.

**lncRNA results**

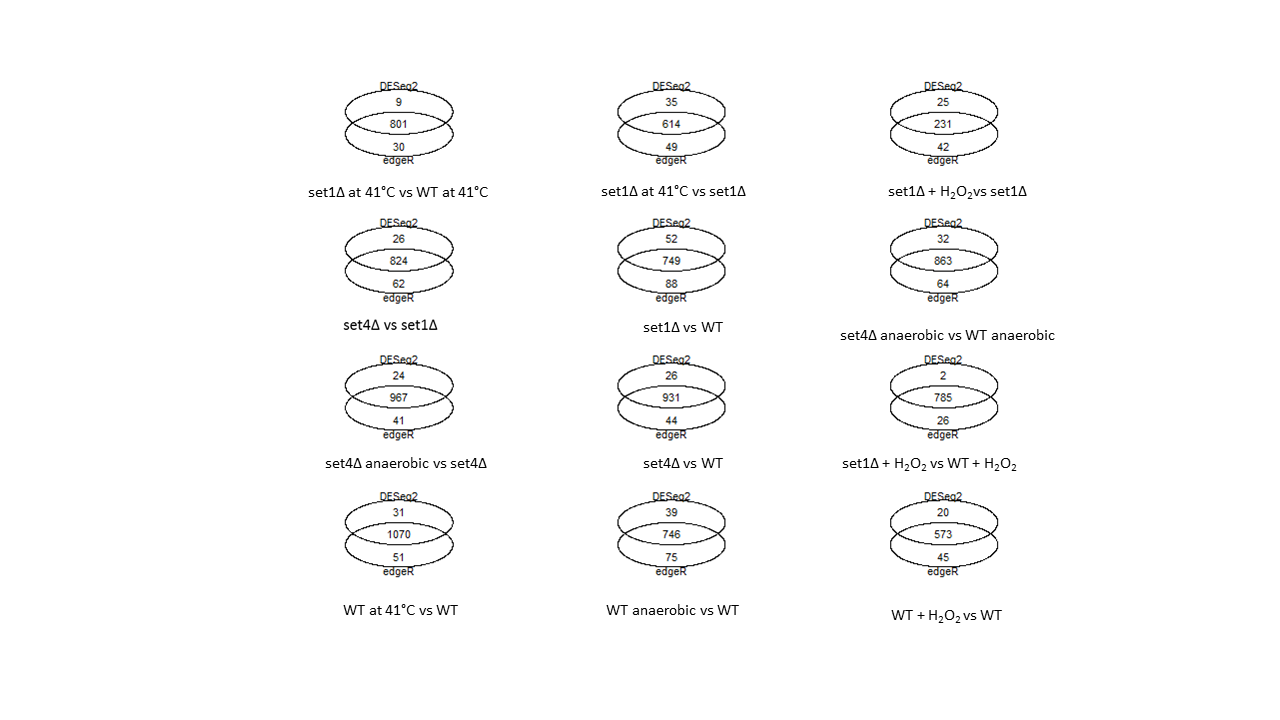
DESeq2 (Love et al., 2014), and edgeR (Robinson et al., 2010; Robinson & Oshlack, 2010 were used to perform a differential expression analysis on lncRNAs. Controlling the FDR at 1%, we found a number of differentially expressed lncRNA when we compared various conditions and genotypes. These differential expression results are summarized in Tables 9 and 10. A total of 2,425 out of 2,523 (~96%) published lncRNAs used had reads aligning. The statistical analysis was run on merged lncRNAs. The reason for this is that there were many published lncRNAs which overlap with one another and are possibly one lncRNA. The merged lncRNAs could be a single CUT/SUT/NUT/XUT if they are not overlapped with each other in the same direction. Or, they could be a merged feature of any types of CUT/SUT/NUT/XUT if they are overlapped. These merged features were re-defined for the purposes of the statistical analysis as merge\_X, with X being an integer. The merge\_X could be intergenic or overlapped with genes. In the latter case, it could associate with one gene or multiple genes. If merge\_X overlaps with multiple genes, there will be multiple rows for merge\_X in the results tables, however the statistics output by DESeq2 would be the same for each of the rows associated with merge\_X. As seen in Figure 19, there is a high degree of overlap between the results of DESeq2 and edgeR.

**Table 9. Differential expression DESeq2 results for lncRNA.** The number of genes found to be differentially expressed by DESeq2 at a 1% false discovery rate are shown, along with the number of lncRNA differentially expressed when an additional 2 fold change cutoff is applied. The number of lncRNA up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 1568 | 785 | 374 | 411 |
| *set1*Δ vs WT | 1562 | 801 | 349 | 452 |
| *set4*Δ vs *set1*Δ | 1719 | 850 | 548 | 302 |
| *set4*Δ vs WT | 1810 | 957 | 550 | 407 |
| *set1*Δ at 41°C vs set1 | 1488 | 649 | 428 | 221 |
| *set4*Δ anaerobic vs *set4*Δ | 1786 | 991 | 435 | 556 |
| *set4*Δ anaerobic vs WT anaerobic | 1623 | 895 | 462 | 433 |
| WT + H2O2 vs WT | 1446 | 593 | 226 | 367 |
| *set1*Δ + H2O2 vs *set1*Δ | 1130 | 256 | 116 | 140 |
| *set1*Δ + H2O2 vs WT + H2O2 | 1585 | 787 | 418 | 369 |
| *set1*Δ at 41°C vs WT at 41°C | 1683 | 810 | 424 | 386 |
| WT at 41°C vs WT | 1719 | 850 | 548 | 302 |

**Table 10. Differential expression edgeR results for lncRNA.** The number of lncRNA found to be differentially expressed by edgeR at a 1% false discovery rate are shown, along with the number of lncRNA differentially expressed when an additional 2 fold change cutoff is applied. The number of lncRNA up-regulated and down regulated are also shown. For each comparison, A vs B, fold change was calculated as condition A/condition B.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison** | **FDR<=0.01** | **# DEGs (FC>2)** | **# up** | **# down** |
| WT anaerobic vs WT | 1601 | 821 | 445 | 376 |
| *set1*Δ vs WT | 1610 | 837 | 435 | 402 |
| *set4*Δ vs *set1*Δ | 1790 | 886 | 534 | 352 |
| *set4*Δ vs WT | 1838 | 975 | 594 | 381 |
| *set1*Δ at 41°C vs set1 | 1538 | 663 | 402 | 261 |
| *set4*Δ anaerobic vs *set4*Δ | 1834 | 1008 | 412 | 596 |
| *set4*Δ anaerobic vs WT anaerobic | 1666 | 927 | 439 | 488 |
| WT + H2O2 vs WT | 1509 | 618 | 267 | 351 |
| *set1*Δ + H2O2 vs *set1*Δ | 1148 | 273 | 93 | 180 |
| *set1*Δ + H2O2 vs WT + H2O2 | 1611 | 811 | 425 | 386 |
| *set1*Δ at 41°C vs WT at 41°C | 1734 | 831 | 449 | 382 |
| WT at 41°C vs WT | 1811 | 1121 | 577 | 544 |



**Figure 12. Venn diagrams of overlap between lncRNA expression results found using edgeR and DESeq2.**

**Pathway analysis and GO enrichment for Differentially Expressed Genes**

DAVID (the Database for Annotation, Visualization, and Integrated Discovery) was used in the annotation of genes and in performing the GO (gene ontology) enrichment analyses. An enrichment analysis was performed for each set of differentially expressed genes, using the intersection from all three packages. GO enrichment results are included in a spreadsheet. Significantly enriched KEGG pathways were also found using DAVID. There are numerous columns for each GO enrichment/KEGG enrichment analysis:

Category: the original database/resource that the terms originated in

Term: enriched terms associated with your gene list

Count: the number of differentially expressed genes associated with the term

Percentage: the percentage of genes involved (involved genes/total genes)

P-value: Modified Fisher Exact P-value, the smaller the value, the more enriched the term

Genes: differentially expressed genes associated with the term

Pop Hits: how many genes in the whole background list have the are associated with this term

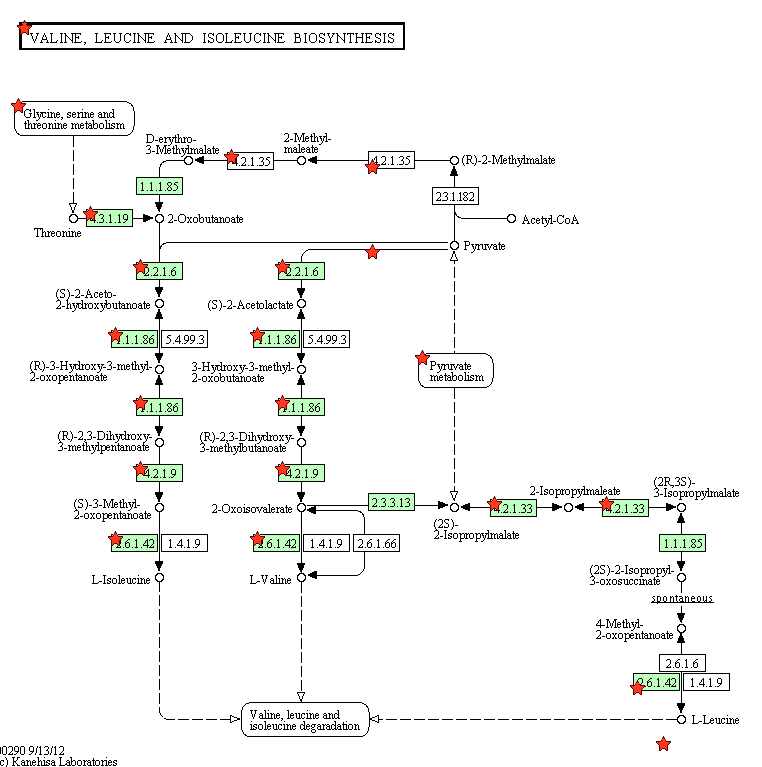
Pop Total: how many genes in the overall population (the yeast genome) are associated with this ontology

Fold Enrichment: numeric to show ratio of two proportions. If 10/100 (10%) of genes are involved in ribosome biogenesis and the background information is 40/40000 (1%) of genes associating with ribosome biogenesis, then 10%/1% = 10 fold enrichment is seen.

Bonferroni: p-value adjusted to correct for multiple testing, corrected using Bonferroni correction

Benjamini: p-value adjusted to correct for multiple testing, corrected using Benjamini’s method

FDR: p-value adjusted to control (false discovery rate)



**Figure 13. Pathway diagram for Valine, leucine, and isolueucine biosynthesis pathway in Set4 anaerobic vs wild type anaerobic comparison.** Differentially expressed genes are starred. Green boxes indicate the yeast-specific pathway. White boxes show genes present in the pathway in other organisms but no present in yeast.

**Summary and Discussion**

Overall, this project was highly fruitful: a number of differentially expressed genes, antisense transcripts, and lncRNAs were identified. There are of course additional analyses we may wish to perform. Pathway diagrams can be made for any of the pathways found to be significantly enriched in the differentially expressed genes. Since the overlap between Cufflinks and DESeq2/edgeR was so poor for antisense transcription, we may also wish to only include the analysis from edgeR/DESeq2 when looking at intersecting genes. More work needs to be done to determine precisely why there is so little overlap between the Cufflinks output and the edgeR/DESeq2 output. Another possibility is to screen with genes usually “on” vs “off” in G2 vs. G1; these genes can be screened out and it can be determined if these are the *SET1* dependent genes. Additionally, we can combine the analyses described in this report with the available complementation data and suppressor screen data which will hopefully enable us to generate even more biologically meaningful conclusions. It is possible that the deletions may lead to changes in read-through transcription rates. FusionCatcher (Nicorici et al., 2014) FusionCatcher was run on reads aligning to the genome, however the program is unable to properly format the necessary databases. The authors have been contacted and are fixing the bug. FusionCatcher will be run to identify likely read-through events as soon as the bug is fixed.

**References**

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166-169. doi: 10.1093/bioinformatics/btu638

Andrews, S. (2010). FastQC.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., . . . Cherry, J. M. (2014). The reference genome sequence of Saccharomyces cerevisiae: then and now. *G3 (Bethesda), 4*(3), 389-398. doi: 10.1534/g3.113.008995

Gordon, A. (2009). FastX-Toolkit. Retrieved from <http://hannonlab.cshl.edu/fastx_toolkit/download.html>

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13. doi: 10.1093/nar/gkn923

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc, 4*(1), 44-57. doi: 10.1038/nprot.2008.211

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol, 14*(4), R36. doi: 10.1186/gb-2013-14-4-r36

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., . . . Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol, 9*(8), e1003118. doi: 10.1371/journal.pcbi.1003118

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi: 10.1186/s13059-014-0550-8

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res, 40*(10), 4288-4297. doi: 10.1093/nar/gks042

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics, 27*(17), 2325-2329. doi: 10.1093/bioinformatics/btr355

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol, 12*(3), R22. doi: 10.1186/gb-2011-12-3-r22

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi: 10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol, 11*(3), R25. doi: 10.1186/gb-2010-11-3-r25

South, P. F., Harmeyer, K. M., Serratore, N. D., & Briggs, S. D. (2013). H3K4 methyltransferase Set1 is involved in maintenance of ergosterol homeostasis and resistance to Brefeldin A. *Proc Natl Acad Sci U S A, 110*(11), E1016-1025. doi: 10.1073/pnas.1215768110

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol, 31*(1), 46-53. doi: 10.1038/nbt.2450

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc, 7*(3), 562-578. doi: 10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol, 28*(5), 511-515. doi: 10.1038/nbt.1621

van Dijk, E. L., Chen, C. L., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., . . . Morillon, A. (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature, 475*(7354), 114-117. doi: 10.1038/nature10118

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., . . . Steinmetz, L. M. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature, 457*(7232), 1033-1037. doi: 10.1038/nature07728