RNA-Seq Analysis of Wildtype Caenorhabditis Elegans Germlines Under Different Temperature Conditions

Chong Li
Marquette University

RNA-SEQ ANALYSIS OF WILDTYPE *CAENORHABDITIS ELEGANS* GERMLINES UNDER DIFFERENT TEMPERATURE CONDITIONS

By

Chong Li

A Thesis Submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science

Milwaukee, Wisconsin

May 2020

ABSTRACT
RNA-SEQ ANALYSIS OF WILDTYPE *CAENORHABDITIS ELEGANS* GERMLINES UNDER DIFFERENT TEMPERATURE CONDITIONS

Chong Li

Marquette University, 2020

Quantitative analysis is very important for researchers to understand the molecular physiology underlying differential gene expression. High-throughput mRNA sequencing (RNA-seq) has become a standard method, which can be used in a wide variety of species and biological conditions to discover new genes and transcripts or measure levels transcript expression. The nematode *Caenorhabditis elegans* is an important model for the study of germ cell biology. For this thesis, RNA-Seq was performed on dissected germlines of *Caenorhabditis elegans* that were grown at either 20°C (ideal conditions) or 27°C (stress conditions) from two wildtype strains: JU1171 (thermotolerant) and LKC34 (thermosensitive). The goals of this research were to uncover four expression patterns that are different between these two strains under two different temperature conditions, which could potentially underlie the phenotypic difference when *Caenorhabditis elegans* are stressed. I performed and compared five different RNA-Seq pipelines, which include *Cuffdiff, DESeq2*, *edgeR, limma*, *DESeq*, starting with 16 raw sequencing *fastq* files, including experimental design, quality control, read alignment, expression quantification, differential gene expression, and enrichment analysis. My research resulted in both differential expression data and analyzed patterns of differentially expressed genes. I also did the enrichment analysis on the functions of genes under each pattern to uncover the different expression patterns between the two strains and two temperatures. From the result, we predict that increased apoptosis at elevated temperatures is protective for fertility. In the end, I discussed the drawbacks in the analysis that can be improved and mentioned additional analysis which can be added to the outcomes in the future.

ACKNOWLEDGEMENTS

Chong Li

I would like to express my deep gratitude to Dr. Lisa Petrella, my thesis advisor, for her professional guidance, valuable assistance and constructive suggestions during the planning and development of this thesis research work since summer 2019. Her willingness to give her time so generously has been very much appreciated.

I would also like to thank my committee members, Dr. Naveen Bansal and Dr. Mehdi Maadooliat, for their useful critiques and enthusiastic encouragement of this research work.

Finally, I would like to offer my special thanks to my parents for their support and encouragement throughout my graduate study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Quantifying genes that are differentially expressed between different strains and conditions in a cell, tissue or organism is a crucial approach for researchers to investigate the molecular mechanisms underlying phenotypes differences (Ji and Sadreyev 2018). RNA-Seq analysis, which is based on next-generation sequencing data, is a recently emerged approach for the analysis of differential gene expression, especially at the whole transcriptome level. Typically, an RNA-Seq workflow includes experimental design, quality control of the raw sequence data, read mapping, expression quantification, differential expression testing, functional interpretation and several biological insights and hypothesis. The workflow described above mainly requires the installation of Unix or Linux and R command-line interfaces, such as RStudio (RStudio Team 2015). In the bioinformatics research area, multiple different methods have been developed to identify differentially expressed genes from various RNA-Seq data, however, there is no consensus exists on which of these methods perform best.

The nematode *Caenorhabditis elegans* is an important laboratory model in biomedical research due to its genetic manipulability, a fully described developmental system, a well-characterized genome, a short and productive life cycle, and also a small body size (Leung et al. 2008). The most suitable growth temperature for *C. elegans* in the lab is about 20°C (Brenner 1974). Between 15°C to 25°C is considered to be the physiological ideal; however, temperature conditions beyond this range are considered stressful and can result in the development or physiology of worms being compromised (Gómez-Orte et al. 2018). Extreme temperature conditions are known to have a primary

negative influence on the physiological parameters of the worm, such as fertility or longevity. While previous studies showed that the standard growth and maintenance temperature for *C.elegans* is 20°C, and that temperatures ranging from 15°C to 25°C are considered physiological conditions, the effect of these conditions on the worm transcriptome had not been well characterized (Gómez-Orte et al. 2018). According to Gómez-Orte et al. (2018), they compared the global gene expression profile for the reference *C. elegans* strain (N2) which was grown at 15°C, 20°C, and 25°C on two different diets, *Escherichia coli* and *Bacillus subtilis.* Their results showed that *C. elegans* undergo significant metabolic and defense response changes when the maintenance temperature fluctuates within the physiological range. Harvey and Viney (2007) state that temperature affects the lifetime fecundity and the reproductive timing of *C. elegans*. They additionally found that there is a genotype by environment interaction, with different wildtype isolates varying in how lifetime fecundity changes with temperature. They found that a reduction in the number of functional sperm was primarily causing the lower lifetime fecundity observed at higher temperatures up to 25°C. According to the Prasad et al. (2010) investigation of the temperature's effects on the fecundity of self-fertilizing nematodes of the species *Caenorhabditis briggsae*, they found that isogenic strains from a Tropical phylogeographic clade have greater lifetime fecundity when reared at extremely high temperatures and lower lifetime fecundity at extremely low temperatures than do strains from a Temperate phylogeographic clade, which is consistent with adaptation to local temperature regimes. Petrella (2014) showed that there were significant temperature, genotype and temperature × genotype effects on fertility of *C. elegans*. For most isolates, 100% of the population maintained fertility from

20°C to 26°C, but there was a steep drop in the percentage of fertile hermaphrodites at 27°C (Petrella 2014). Also, in the Poullet et al. (2015) paper, they found that temperature variation modulates spermatogenesis, oogenesis and germ cell progenitor pools, which is consistent with evolutionary variation in upper thermal limits of hermaphrodite fertility. High temperature significantly perturbs oogenesis, germline integrity, and mitosis–meiosis progression, even though defective sperm function is a major contributor to heat-induced fertility breakdown (Poullet et al. 2015). These studies showed that temperature will influence the lifetime fertility and functional differential gene expression of *C. elegans*. However, there has not been a study of the changes in gene expression that may underlie the differences between the different wild type strains *C. elegans* strains in fertility under different temperature conditions. In my thesis, RNA-Seq analysis was applied to further define the biological processes that change under stress conditions.

In this thesis, the most frequently used RNA-sequencing methods were applied and compared on dissected germlines of *Caenorhabditis elegans* to ultimately come to the lists of differentially expressed genes of four potential expression patterns between the two strains under the two temperature conditions. It was found that there were significantly genotype × environment differences such that some strains are much more thermal tolerant, and others are much more thermal sensitive. Two wild type strains were used in my thesis research: JU1171, which is thermotolerant, and LKC34, which is thermosensitive. There are a significantly higher number of JU1171 fertile hermaphrodites and higher brood size compared to LKC34 at 27°C (Petrella 2014). For each genotype under both 20°C and 27°C, there were four biological replicates dissected from germline and sent for RNA-sequencing. Thus, a totally16 raw sequencing *fastq* files

data were analyzed. The goals of this thesis research were to find the molecular differences which could potentially underlie the reasons the JU1171 strain maintained a higher level of fertility at 27°C compared to LKC34. I applied and compared several RNA-Seq analysis pipelines, *Cuffdiff, DESeq2*, *edgeR*, *limma* and *DESeq*, to define four expression patterns and did additional analysis of the types of genes within these four pattern groups to allow for further definition of the biological processes that change under stress conditions, such as enrichment analysis of functional gene categories and Gene Ontology analysis. Particularly, two different read alignment tools, *TopHat2* and *HISAT2*, were compared and chosen by both the alignment rate and mapping time. Two read counting tools, *HTSeq-Count* and *featureCounts*, were both used for different purposes of downstream analyses and visualization. The detailed description of the parameter settings of the methods and software are introduced in the Methods and Materials section.

To summarize, my thesis highlights the main *C. elegans* transcriptomic response differences when two different strains (thermotolerant and thermosensitive) are cultivated at two different temperature conditions (ideal condition and stress condition). In particular, four potential expression patterns were mainly analyzed and several similarities and dissimilarities within the four RNA-Seq analysis methods were mainly compared. Gene expression differences reflected the different physiologically mechanisms and phenotypic of worms in response to a higher temperature. Based on the analysis of pattern 1, we have a prediction that increased apoptosis at elevated temperatures is protective for fertility. I end with some discussions and improvements for further research.

2. MATERIALS AND METHODS

2.1 *C. elegans* Strains and Temperature Treatments

Two wild type strains were used in my thesis research: JU1171, which is thermotolerant, and LKC34, which is thermosensitive (Petrella 2014). Worms were cultured using standard conditions (Brenner 1974) at 20°C unless otherwise noted. For the 20°C experiment, worms were continuously maintained at 20°C. For the 27°C experiment, P0 hermaphrodites were upshifted from 20°C to 27°C at the L4 stage and F1 animals maintained continuously at 27°C.

2.2 Gonad Dissection and RNA Isolation

Gonads were dissected from young adult animals approximately 24 hours after the L4 stage in 1X egg buffer (25 mM HEPES, 120 mM NaCl, 2 mM MgCl2, 2 mM CaCl2, 50 mM KCl). Gonads were cut between the last oocyte and the spermatheca and were placed directly into 100µl Trizol (Invitrogen, cat#15596026). Four biological replicates were done per genotype per temperature. Between 52 to 108 gonads were used per RNA sample isolated. Each sample was ground with a pestle and then 200µl Trizol added. Then total RNA was isolated using the Zymo Direct-Zol miniprep kit using the manufacturer's instructions including on-column DNase I digestion. Elution was done using 25µl DNase free water. Total RNA was stored frozen at ≤ -80°C until sequencing was done.

2.3 Library Preparation and Sequencing

The University of Wisconsin-Madison Biotechnology Center Gene Expression Center prepared libraries for each sample. RNA samples were thawed on ice and each sample assayed on the NanoDrop2000 (quantification) and Agilent RNA PicoChip (quality). cDNA sequencing libraries from four biological replicates were prepared from total RNA from each strain (JU1171 and LKC34) of each condition (20°C and 27°C) by following the standard protocol from Illumina Stranded TruSeq RNA Library Preparation Kit v2 to poly-A enrichment and fragment mRNA. Raw sequence reads were obtained from the Illumina HiSeq2000. Library concentration was assessed, and each library assayed in a singlet with a 1:100 dilution before high throughput sequencing.

2.4 Quality-control Checkpoints

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to perform quality control analyses on Illumina reads (Conesa et al. 2016) from the command line or as a graphical application on the *fastq* file from the sequencer. Software tools such as *Trim_Galore* can be used to discard low-quality reads, trim adaptor sequences, and eliminate poor-quality bases. *Trim_Galore* was run from the command line and was a wrapper around another program called *Cutadapt* using default options and adapters that were detected were removed. *Trim_Galore* produced a trimming report which I can have a look through to see details of any trimming that was carried out.

2.5 Read Alignment

Raw sequence reads were obtained from the Illumina HiSeq2000 and processed as single-end reads using two different pipelines. First, *TopHat2* was used to align reads to the *C. elegans* reference genome (WBcel235.96.fa) and gene annotations (WBcel235.96.gtf) in NCBI and WormBase WS271. *TopHat2* (v2.1.1) used *Bowtie* as an alignment engine and broke up reads that *Bowtie* cannot align on its own into smaller pieces called segments. By processing each initially unmappable read, *TopHat2* built up an index of splice sites in the transcriptome (Trapnell et al. 2012). Second, *HISAT2* was used to perform the read alignment, which was designed as a successor to *TopHat* and *TopHat2* (Pertea et al. 2016). Default parameters were used in most instances, with the following exceptions: *--read-mismatches*; *--read-gap-length* and *--read-edit-dist* arguments were specified in the *TopHat2* protocol; *--mp* and *--rdg* arguments were used in the *HISAT2* command which I will discuss more in the results part. Finally, *HISAT2* output a SAM file, then *Samtools* was used to compress the raw SAM format output into the more compact sorted BAM format.

2.6 Differential Expression Testing

2.6.1 Counting reads in feature with *HTSeq-Count* and *featureCounts*

After running *Samtools,* the resulting sorted BAM files were provided to *HTSeq*, which is a Python package that calculates the number of mapped reads to each gene. *HTSeq-Count* files are available in a tab-delimited format with one Ensembl gene ID column and one mapped reads column for each gene (Anders and Huber 2014). In order to get a normalized read count, specifically FPKM (fragment/reads per kilobase of

transcript per million mapped reads) value, which is not an output, I chose to use *featureCounts* (Liao et al. 2014). *featureCounts* is another reliable reads counting software and performs about 20 times quicker than *HTSeq-Count* (Yang et al. 2014). The output of the *featureCounts* program includes a count table and a summary of counting results which are saved into two-delimited files. One of the files is the table that includes the read counts and contains 'Geneid', 'Chr', 'Start', 'End', 'Strand' and 'Length' five annotation columns. The other one contains the summary of counting results which is named of reads counts file added with '. summary' (Liao et al. 2019).

2.6.2 Analyzing RNA-Seq data with *DESeq2*

The package *DEseq2* provides methods to test for differential gene expression (Love et al. 2014). As input, the *DESeq2* package requires count data to be input in the form of a matrix of integer values (Anders et al. 2010). The count data output from the *HTSeq-Count* was used. The value in the *x-th* row and the *y-th* column of the matrix tells us how many reads can be assigned to gene x in sample y. The values in the matrix should be un-normalized counts or estimated counts of sequencing reads for single-end RNA-Seq or fragments for paired-end RNA-Seq and the *DESeq2* model internally corrects for library size, so transformed or normalized values such as counts scaled by library size should not be used as input. Then the function *DESeqDataSetFromMatrix* was used. A minimal pre-filtering was performed to keep only rows that have at least 1 read in total. The function results were used to generate the results table with log2 fold changes, *p*-values and adjusted *p*-values that using the "BH" method of Benjamini and Hochberg that controls the false discovery rate (Benjamini and Hochberg 1995). Then the adjusted *p*-value (*q*-

value) of 0.05 and fold changes of 1.5 were used as the criteria to get the differential expressed gene lists. Since the output of the *featureCounts* gives us the length of each gene, I performed an *fpkm()* function in *DESeq2* with the length value and then used the *ggplot2* package to get several scatter plots among the different replicates in the same genotype under the same condition and also output the scatter plots among the different replicates in the same genotype under different condition.

2.6.3 Analyzing RNA-Seq data with *edgeR*

edgeR (Robinson et al. 2010) is the short name of Empirical Analysis of Digital Gene Expression Data in R. It is a package for the differential expression analysis of digital gene expression data. As with *DESeq2*, *edgeR* also works on a table of integer reads counts, which I got from *HTSeq-Count*. After reading the counts tables, *edgeR* stores data in a simple list-based data object called a DGEList and then added a grouping factor that includes our 16 sample names for short. CPM (count-per-million) was used to filter out lowly expressed genes. The function *calcNormFactors* was used to do the normalization by finding a set of scaling factors for the library sizes (Robinson et al. 2010). The function *model.matrix* was used to construct the design matrix. *edgeR* uses a special method called quantile-adjusted conditional maximum likelihood (qCML) for experiments with a single factor. The qCML common dispersion was estimated using the *estimateDisp* function on the DGEList object (Chen et al. 2014). Since the quasi-likelihood (QL) F-test is preferred as it reflects the uncertainty in estimating the dispersion for each gene, functions *glmQLFit* and *glmQLFTest* were used to perform the QL dispersion estimation and hypothesis testing (Lun et al. 2016).

2.6.4 Analyzing RNA-Seq data with *limma*

limma, which is Linear Models for Microarray and RNA-Seq Data for short, is a package for the analysis of gene expression data arising from microarray or RNA-Seq technologies (Ritchie et al. 2015). I used the same filtered and normalized gene lists as the *edgeR* method used. Then, read counts were converted to log2-counts-per-million (logCPM) in the *limma*-trend approach using *edgeR*'s *cpm* function (Law et al. 2014). Estimated the fold changes and standard error by fitting a linear model for each gene using the *lmFit* function and applied empirical Bayes smoothing to the standard errors by using the *eBayes* function (Phipson et al. 2016). The function *topTable* was used to display the results of the top genes. In the end, I used the same criteria of fold changes and *q*-value as the previous two methods used to filter the higher expressed gene in each comparable group by using the subset function.

2.6.5 Analyzing RNA-Seq data with *DESeq*

In addition to *DESeq2*, *edgeR*, and *limma*, which are the most popular three methods of deferential expression analysis, many research papers also applied *DESeq* as their main method. In this thesis, the same RNA-sequencing and analysis protocols were run as described in Campbell and Updike (2015): *TopHat2* (v.2.0.8b) was run to map the reads, *HTSeq* was used to count reads number per gene per sample and *DESeq* was then run with default parameters to test the differential gene expression. As previously described, *HISAT2* was eventually chosen for use in my thesis after the comparison of several aspects between *TopHat2*. Thus, *TopHat2* was also run combined with the *DESeq*

to test if there was any big difference between these two protocols. The detailed description of this comparison was shown in the following results section.

2.7 Analysis of Gene Set

2.7.1 Pattern analysis

In this thesis, four different patterns were defined based on differential gene expression results from *DESeq2*. Pattern 1 was defined as those genes that were up-regulated in JU1171 at 27°C compared to JU1171 at 20°C but were not up-regulated in LKC34 at 27°C compared to LKC34 at 20°C. Pattern 2 was defined as those genes that were down-regulated in JU1171 at 27°C compared to JU1171 at 20°C but were not down-regulated in LKC34 at 27°C compared to LKC34 at 20°C. Pattern 3 was defined as genes that were expressed higher in JU1171 compared to LKC34 at both temperatures. Pattern 4 was defined as genes that were expressed higher in LKC34 compared to JU1171 at both temperatures. I counted the number of genes in each pattern by filtering the TURE of FALSE values for the corresponding columns and got the gene ID lists for each pattern to do the following enrichment analysis.

2.7.2 Enrichment analysis

2.7.2.1 Gene set analysis

The hypergeometric test is a statistical test which uses the hypergeometric distribution to calculate the statistical significance to identify which sub-populations are over-represented or under-represented in a specific sample (Rivals et al. 2017). In my thesis research, the hypergeometric test was used to ask if genes normally expressed in

the following gene sets: Germline enriched genes (2464 genes) (Reinke et al. 2004), Germline enriched gender neutral (908 genes) (Reinke et al., 2004), Soma enriched genes (327 genes) (Reinke et al. 2004), Neuron enriched genes (1324 genes) (Watson et al. 2008), Spermatogenesis enriched genes (754 genes) (Reinke et al. 2004), Spermatogenesis enriched genes (2221 genes) (Ortiz et al. 2014), List of genes encoding spermatogenesis proteins (103 genes) (Chu et al. 2006), Oocyte enriched genes (809 genes) (Reinke et al. 2004), Oocyte enriched genes (1512 genes) (Ortiz et al. 2014), List of the significantly up-regulated genes altered in response to the "hsf-1(+);+Heat-Shock vs control" condition (673 genes) (Brunquell et al. 2016) and List of the significantly down-regulated genes altered in response to the "hsf-1(+);+Heat-Shock vs control" condition (357 genes) (Brunquell et al. 2016), were enriched for or depleted from genes that are represented by the four defined patterns. The cut-off *p*-value was set as 0.01.

2.7.2.2 Gene ontology (GO) analysis

Two different enrichment analyses were used for Gene Ontology Analysis: PANTHER Overrepresentation Test (http://geneontology.org/) and gProfiler (https://biit.cs.ut.ee/gprofiler/gost). The genes from each of the four defined patterns were uploaded to both the enrichment analysis tools and significantly enriched gene ontology terms, their associated *q*-value and other statistics values were obtained along.

3. RESULTS

My thesis was done using data 16 RNA-Seq datasets. These represent the sequencing of mRNA from dissected germlines of two different wild type strains of *C. elegans*, JU1171 and LKC34 with four biological replicates from two temperature treatments 20°C and 27°C. The goal of analyzing these data sets was to find the molecular differences which could potentially underlie the higher level of fertility in JU1171 than LKC34 at the higher temperature. By checking the *FastQC* Reports for each of the samples, we found that all of our data had high quality and there was no need for trimming before alignment, which allowed us to directly use the data to do the following analysis.

3.1 *HISAT2* Performs Better than *TopHat2* for Read Alignment

RNA-Seq analysis begins by mapping reads against a reference genome to identify their genomic position (Trapnell et al. 2012). It is a major step in the analysis pipelines for RNA-seq. Sequence alignment itself is a classic problem in computer science and appears frequently in the bioinformatics area. Therefore, many read alignment programs have been developed within the last few years. One of the most popular RNA-Seq mappers, *TopHat*, follows a two-step strategy in which unspliced reads are first mapped to locate exons, then unmapped reads are split and aligned independently to identify exon junctions (Conesa et al. 2016). *TopHat2* (v2.1.1), which uses *Bowtie* as an alignment engine and breaks up reads that *Bowtie* cannot align on its own into smaller pieces called

segments. By processing each initially unmappable read, *TopHat2* can build up an index of splice sites in the transcriptome (Trapnell et al. 2012).

RNA-Seq mappers need to solve an additional problem that is not encountered in DNA-only alignment: many RNA-Seq reads will span introns. *HISAT2* uses two types of indexes for alignment: a global whole-genome index and tens of thousands of small local indexes. Both these two types of the index are constructed using the same BWT/FM index as *Bowtie2*, and the *HISAT2* system even uses some of the *Bowtie2* code. Because *HISAT2* uses these efficient data structures and algorithms, it generates spliced alignments several times faster than *Bowtie* and BWA while using only about twice as much memory (Pertea et al. 2016). *HISAT2* was designed as a successor to *TopHat* and *TopHat2*, it runs about 50 times faster than *TopHat2* and gives higher alignment rate results (Table1 and Table2).

We wanted to test the optimal parameters for each of these two methods and chose to use the second replicate of LKC34 at 27°C as a test sample. Default parameters were kept in most instances while I focused modulating two parameters of *HISAT2*: "*--mp*" and "*--rdg*", and three parameters of *TopHat2*: "*--read-mismatches*", "*--read-gap-length*" and "*--read-edit-dist*". All of these parameters have the primary influence on the final alignment rate, where a higher alignment rate is generally considered better. "*--mp*" represents maximum (mx) and minimum (mn) mismatch penalties. "*--rdg*" represents the read gap open and extend penalties. "*--read-mismatches*" represents that final read alignment having more than these many mismatches are discarded. "*--read-gap-length*" represents that final read alignment having more than these many total lengths of gaps are discarded. "*--read-edit-dist*" represents those final read alignments having more than

these many edit distance are discarded. For *HISAT2*, I set 10 different groups of values for the two parameters to compare each of its alignment rates and also set 19 different groups of values for the three parameters of *TopHat2* to find the highest final mapped rate and shortest overall time cost (Table 1 and Table 2). To conclude, for *HISAT2*, lower mismatch penalties and gap penalties can get higher alignment rates. For *ToHat2*, allowing higher mismatches and gaps will get higher mapped rates and after 50 mismatches, gaps and edit-length, the mapped rate will keep as 98.9% which is the highest mapped rate we can get so far. Eventually, we chose to use *HISAT2* protocol where "--*mp* 5,2" and "--*rdg* 4,3" were specified in the command to align all of our 16 sequences data because these parameters setting gave us the highest alignment rate compared with other (we do not want the minimum and extend penalties to be 1, which is too small). We also added "--*dta-cufflinks*" option to report alignments tailored specifically for *Cufflinks*. Ideally, all of the samples had exactly one-time alignment rates that were higher than 94%, which represents that more than 94% reads were uniquely aligned. For the 16 RNA-seq datasets there were no overall alignment rates that were below 96% (Table 3).

Table 1. Parameter setting comparisons for *HISAT2*

	--mp	--rdg	Alignment rate	Aligned exactly one time
1	6,2 (default)	5,3 (default)	97.10%	95.32%
2	7,2	6,3	97.03%	95.26%
3	**5,2**	**4,3**	**97.16%**	**95.38%**
4	3,1	2,1	97.38%	95.60%
5	2,1	2,1	97.73%	95.92%
6	2,1	3,1	97.73%	95.92%
7	2,1	5,1	97.73%	95.91%
8	2,1	9,1	97.71%	95.90%
9	9,5	8,4	96.71%	94.96%
10	9,1	2,1	96.96%	95.20%

"--mp" is maximum (mx) and minimum (mn) mismatch penalties. "--rdg" is the read gap open and extend penalties. "Alignment rate" represents the overall alignment rate. The numbers in bold are the best parameter values we confirmed.

Table 2. Parameter setting comparisons for *TopHat2*

	Mismatch	Gap	Edit-length	Mapped rate	Time
1	1	1	2 (default)	89.80%	0:44:07
2	2 (default)	2 (default)	2 (default)	92.60%	0:32:45
3	3	3	3	93.90%	0:36:24
4	5	5	5	95.30%	0:31:54
5	6	6	6	87.50%	0:39:21
6	7	7	7	95.90%	0:46:44
7	8	8	8	96.10%	0:45:06
8	9	9	9	96.40%	0:45:35
9	10	10	10	97.00%	0:43:57
10	20	20	20	98.20%	0:37:38
11	30	30	30	98.60%	0:38:19
12	40	40	40	98.70%	0:41:17
13	45	45	45	98.70%	0:43:38
14	47	47	47	98.70%	0:43:21
15	48	48	48	98.70%	1:05:28
16	49	49	49	98.70%	0:44:37
17	50	50	50	98.80%	0:44:54
18	100	100	100	98.80%	0:46:07
19	200	200	200	98.80%	0:49:08

"Mismatch" represents that final read alignment having more than these many mismatches are discarded. "Gap" represents that final read alignment having more than these many total lengths of gaps are discarded. "Edit-length" represents those final read alignments having more than these many edit distance are discarded. "Mapped rates" represents the overall alignment rate. "Time" represents that overall time cost.

Table 3. Alignment results for *HISAT2*

Sequences	Total reads	Aligned 0 times	Aligned exactly 1 time	Aligned >1 time	Overall alignment rate
JU1171_rep1_27	20287168	556150 (2.74%)	19372823 (95.49%)	358195 (1.77%)	97.26%
JU1171_rep2_27	18686161	633756 (3.39%)	17738620 (94.93%)	313785 (1.68%)	96.61%
JU1171_rep3_27	18597563	470902 (2.53%)	17823398 (95.84%)	303263 (1.63%)	97.47%
JU1171_rep4_27	20220260	502758 (2.49%)	19399045 (95.94%)	318457 (1.57%)	97.51%
JU1171_rep1_20	17144213	626050 (3.65%)	16202150 (94.51%)	316013 (1.84%)	96.35%
JU1171_rep2_20	20131550	492881 (2.45%)	19334884 (96.04%)	303785 (1.51%)	97.55%
JU1171_rep3_20	18359690	489360 (2.67%)	17561932 (95.65%)	308398 (1.68%)	97.33%
JU1171_rep4_20	17328898	529413 (3.06%)	16547190 (95.49%)	252295 (1.46%)	96.94%
LKC34_rep1_27	19235721	527068 (2.74%)	18391023 (95.61%)	317630 (1.65%)	97.26%
LKC34_rep2_27	16726217	489605 (2.93%)	15939935 (95.30%)	296677 (1.77%)	97.07%
LKC34_rep3_27	17818884	524827 (2.95%)	16995287 (95.38%)	298770 (1.68%)	97.05%
LKC34_rep4_27	16684730	458508 (2.75%)	15953915 (95.62%)	272307 (1.63%)	97.25%
LKC34_rep1_20	18604822	505709 (2.72%)	17795223 (95.65%)	303890 (1.63%)	97.28%
LKC34_rep2_20	17614169	623736 (3.54%)	16684200 (94.72%)	306233 (1.74%)	96.46%
LKC34_rep3_20	20167606	509309 (2.53%)	19359382 (95.99%)	298915 (1.48%)	97.47%
LKC34_rep4_20	18686760	529277 (2.83%)	17852202 (95.53%)	305281 (1.63%)	97.17%

"Sequences" represents each sample of our 16 RNA-Seq data and the total reads corresponding to each of the samples. "Aligned 0 times" indicates the reads that were filed to aligned. "Aligned exactly 1 time" represents uniquely aligned reads while "Aligned >1 time" represents multi-mapped reads. The last column showed the over alignment rate for each sample.

3.2 Scatter Plots of the Gene Expression of Replicates among Samples

We wanted to detect any abnormalities present in our data by investigating the distribution of read counts for each sample and replicate. Ideally, all of the samples or replicates would display similar overall distributions. Scatter plots can be used to visualize the comparison of expression levels between two samples or two treatment groups, where each dot represents a single gene. Scatter plots usually use normalized expression values rather than raw counts to compare the expression levels. Normalized expression values are often in the form of FPKM (fragment/reads per kilobase of transcript per million mapped reads) (McDermaid et al. 2019).

In this thesis, for each strain under each condition, there were four biological replicates that analyzed. In total, 24 scatter plots were created to compare the gene expression of each replicate among the same genotype under the same temperature condition (Figure 1, Appendix 1) and also 32 scatter plots were created to compare the gene expression of each replicate of the same genotype under the different temperature condition (Figure 2, Appendix 2). All of these scatter plots were generated by applying *ggplot*() function in R (R core team 2019), which is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

All of the plots (Appendix 1 and Appendix 2) demonstrated that whatever under the same condition or not, each replicate within the same genotype had highly similar expression patterns across all genes, where it can be seen that a closer clustering of all dots lies exactly at the diagonal line. To compare figure 1 and 2, all of the plots in figure 1 looks tighter than the plots in figure 2. This is consistent with our hypothesis that under

the same temperature condition, the gene expression was highly similar, while there were some differences between the gene expression level when they were under the different temperature conditions. To conclude, there is not much of an effect of the different replicates within the same genotype under the same or different conditions.

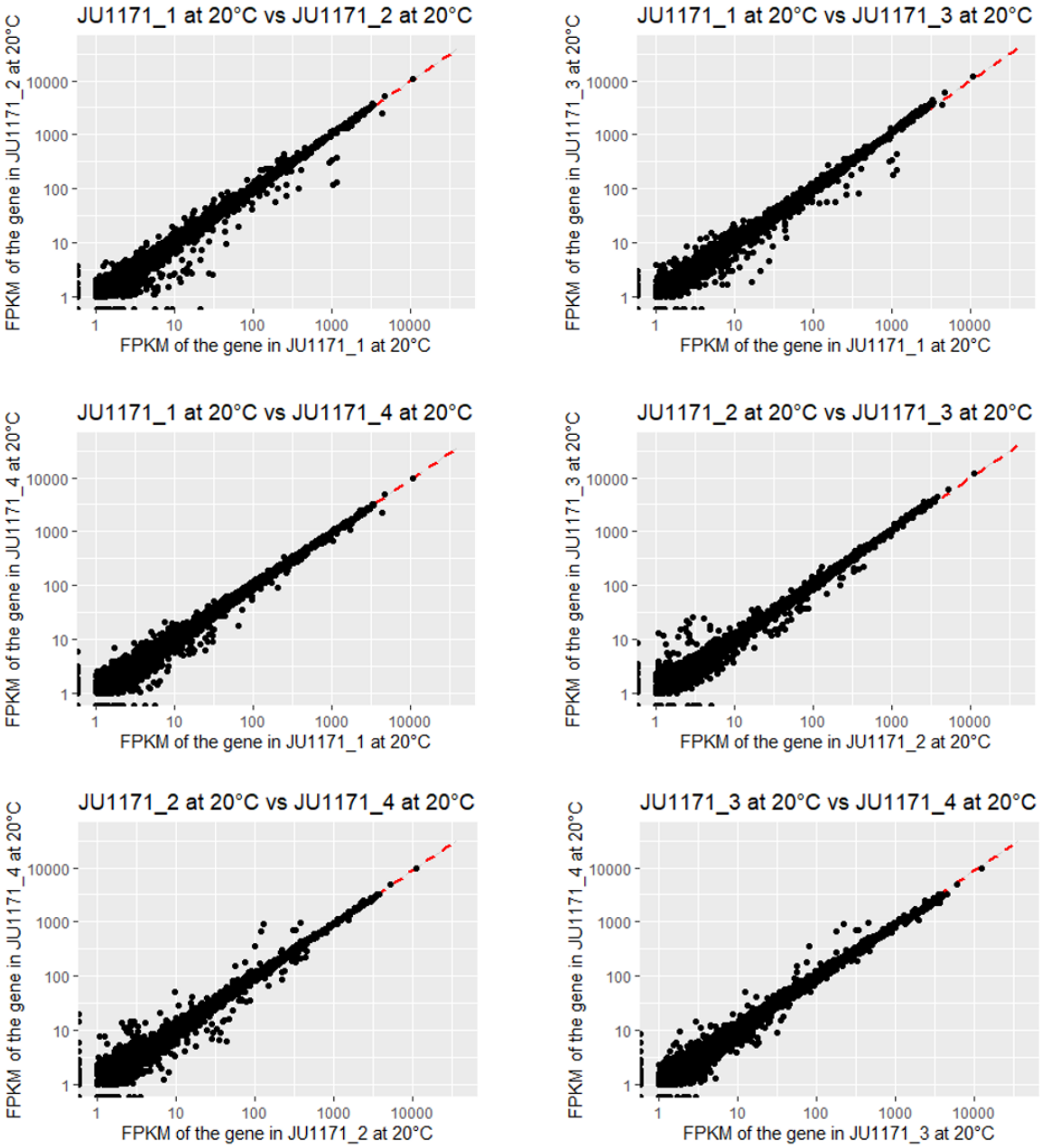
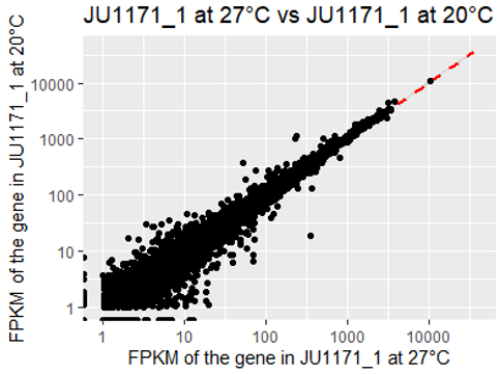
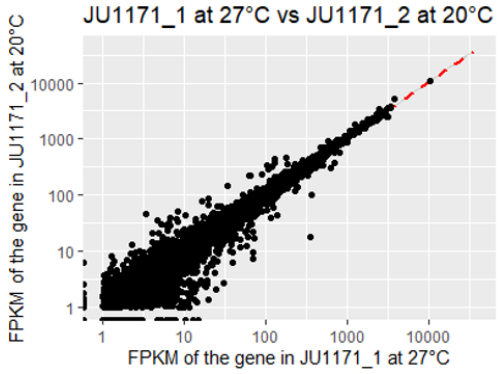
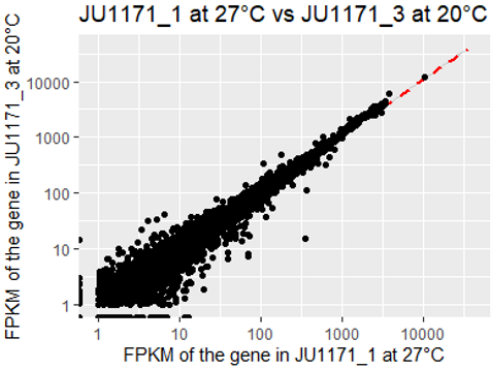
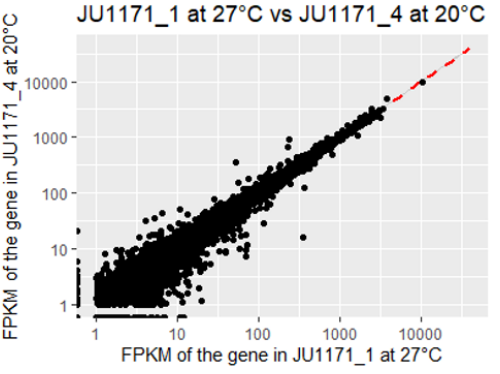


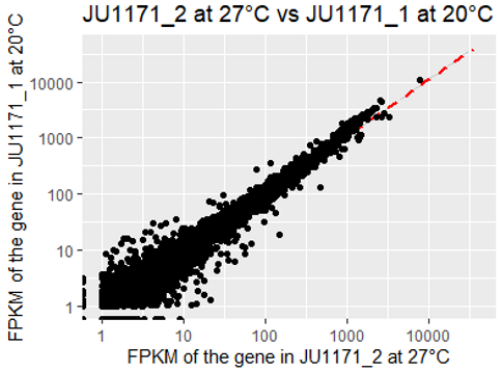
Figure 1. Scatter plots were created to compare the gene expression of each replicate for JU1171 at 20°C (full plots can be found in Appendix 1). The red dashed lines are the regression lines. X-axis and Y-axis are the FPKM values of the gene in the two replicates respectively. Axes were rendered on the log10 scale.

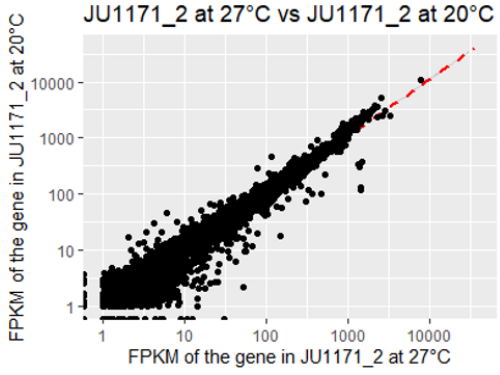

JU1171_1 at 27°C vs JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

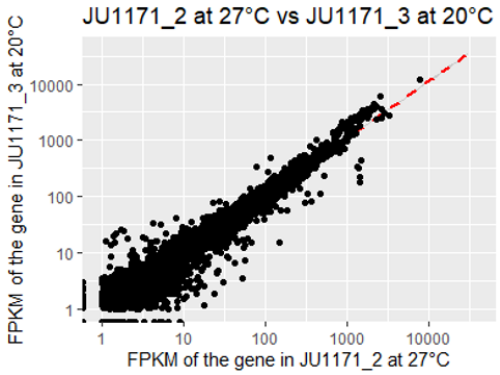

JU1171_1 at 27°C vs JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

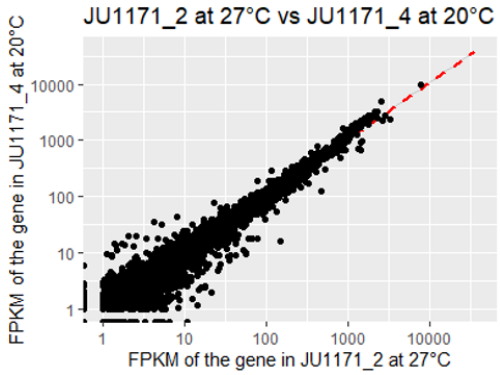

JU1171_1 at 27°C vs JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000


JU1171_1 at 27°C vs JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000


JU1171_2 at 27°C vs JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000


JU1171_2 at 27°C vs JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000


JU1171_2 at 27°C vs JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000


JU1171_2 at 27°C vs JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

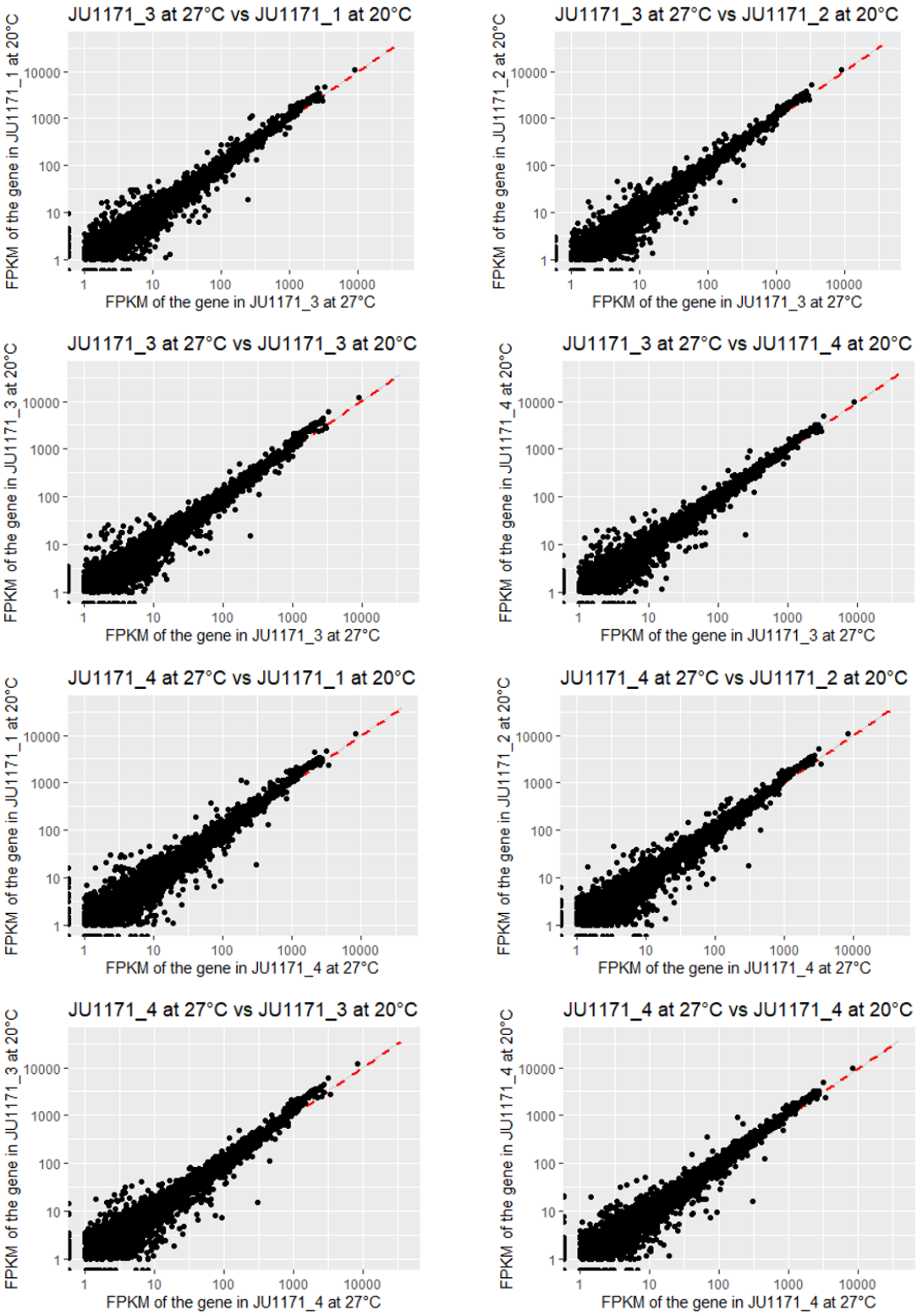


Figure 2. Scatter plots were created to compare the gene expression of each replicate among the JU1171 at 20°C versus 27°C (full plots can be found in Appendix 2). The red dashed lines are the regression lines. X-axis and Y-axis are the FPKM values of the gene in the two replicates respectively. Axes were rendered on the log10 scale.

3.3 Transcript Assembly and FPKM Normalization with *Cufflinks* and *Cuffmerge*

Accurately quantifying the expression level of a gene from RNA-Seq reads requires accurately identifying which isoform of a given gene produced each read, which depends on knowing all of the isoforms of that gene. Attempting to quantify gene and transcript expression by using an incomplete or incorrect transcriptome annotation leads to inaccurate expression values. *Cufflinks* assembles individual transcripts from RNA-Seq reads that have been aligned to the genome. Since a sample might contain reads from multiple splice variants for a given gene, *Cufflinks* must be able to infer the splicing structure of each gene. However, sometimes the gene has multiple alternative splicing events so that there may be many possible reconstructions of the gene model to explain the sequencing data. The truth is that usually, it is not obvious how many splice variants of the gene may be present.

The most common application of RNA-Seq is to estimate gene and transcript expression. This application is primarily based on the number of reads that map to each transcript sequence. The simplest approach to quantification is to aggregate raw counts of mapped reads. Raw read counts alone are not sufficient to compare expression levels among samples because these values are affected by factors such as transcript length, the total number of reads, and sequencing biases. The measure FPKM (Fragments Per Kilobase of transcript per Million mapped reads) is a within-sample normalization method that will remove the feature-length and library-size effects (Conesa et al. 2016).

I used the *Cufflinks* to assemble for each sample with the default parameters to run all the resulting 16 different *.bam* files that were transformed from 16 *.sam* files by using the *samtools*. *Cufflinks* uses FPKM values to report transcript abundances (16

gene.fpkm_trancking files), which reflect the normalization of our RNA-Seq data for depth (average number of reads from a sample that align to the reference genome) and gene length. The counts need to be normalized for the length of a gene to compare the expression levels between genes due to the reason that genes have different lengths (Amrit et al. 2017). In this thesis, there were four biological replicates for each genotype under each condition that analyzed. The four replicates were combined together as one genotype by condition set and these four sets were compared in six ways (Figure 3). For each of the six comparisons, *.txt* files were created, where each of the files listed eight assembly files for two comparisons between groups made up of four replicates. All the assemblies were then merged together along with the reference genome by using the next tool *Cuffmerge* to generate one final assembly containing all transcripts identified across all samples for each of the six comparisons.

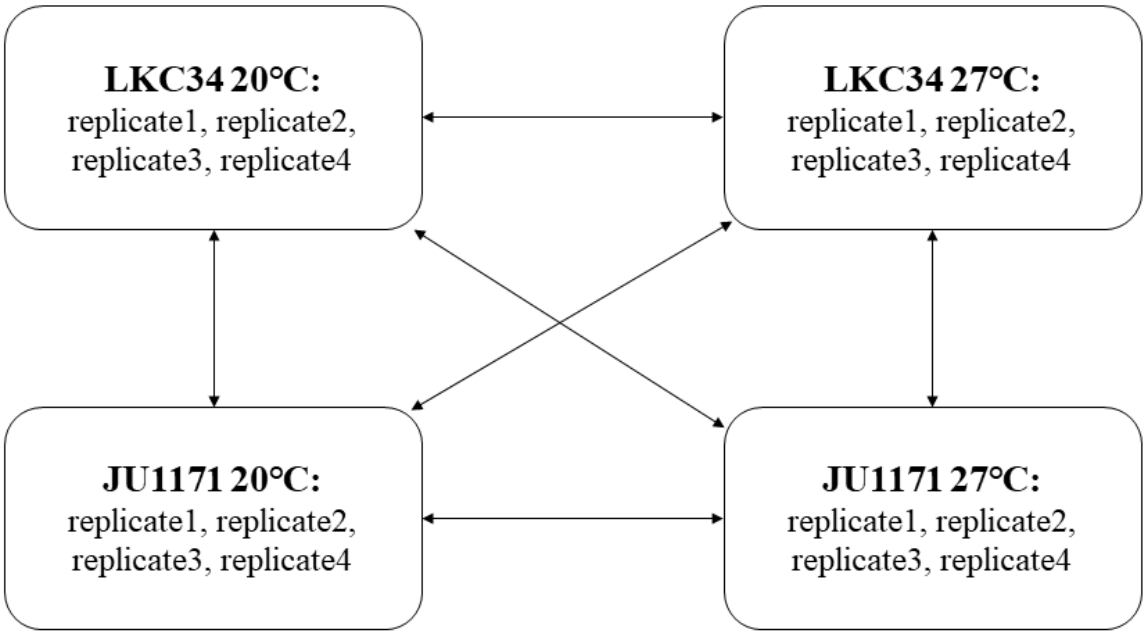


Figure 3. Six comparisons between groups each made up of four biological replicates: LKC34 at 20°C vs. LKC34 at 27°C, JU1171 at 20°C vs. JU1171 at 27°C, LKC34 at 27°C vs. JU1171 at 27°C, LKC34 at 20°C vs. JU1171 at 20°C, LKC34 at 20°C vs. JU1171 at 27°C, LKC34 at 27°C vs. JU1171 at 20°C.

3.4 Differential Expression Testing with *Cuffdiff* and Visualization with R

Cufflinks includes a separate program, *Cuffdiff*, which calculates expression in two or more samples and tests the statistical significance of each observed change in expression between them. With multiple replicates, *Cuffdiff* can learn how read counts vary for each gene across the replicates and used these variance estimates to calculate the significance of observed changes in expression. The reads and the merged assembly were fed to *Cuffdiff*. *Cuffdiff* allows people to supply multiple technical or biological replicate sequencing libraries per condition and provides analyses of differential expression and regulation at the gene and transcript levels.

But browsing these files is not very easy and straightforward, so the *CummeRbund* package for R/Bioconductor was used, which can help people manage, visualize and integrate all of the data produced by a *Cuffdiff* analysis (Trapnell 2012). We can create publication-ready plots with a single command (Trapnell et al. 2012). R is a programming language and free software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis (Fox 2005).

I ran *Cuffdiff* with the default parameters for the six comparisons between groups. It reported many output files containing the results of its differential analysis of the samples which were reported in a set of tab-delimited text files that can be opened with any spreadsheet application, such as Microsoft Excel (Trapnell et al. 2012). I mainly used *gene_exp.diff* which contained familiar statistics such as fold change in log2 scale, *p*-values, *q*-values (FDR adjusted *p*-value) and gene-related and transcript-related attributes such as common name and location in the genome. I generated six scatter plots to

compare the expression of each gene for each comparison (Figure 4) and also six volcano plots were created to inspect differentially expressed genes (Figure 5).

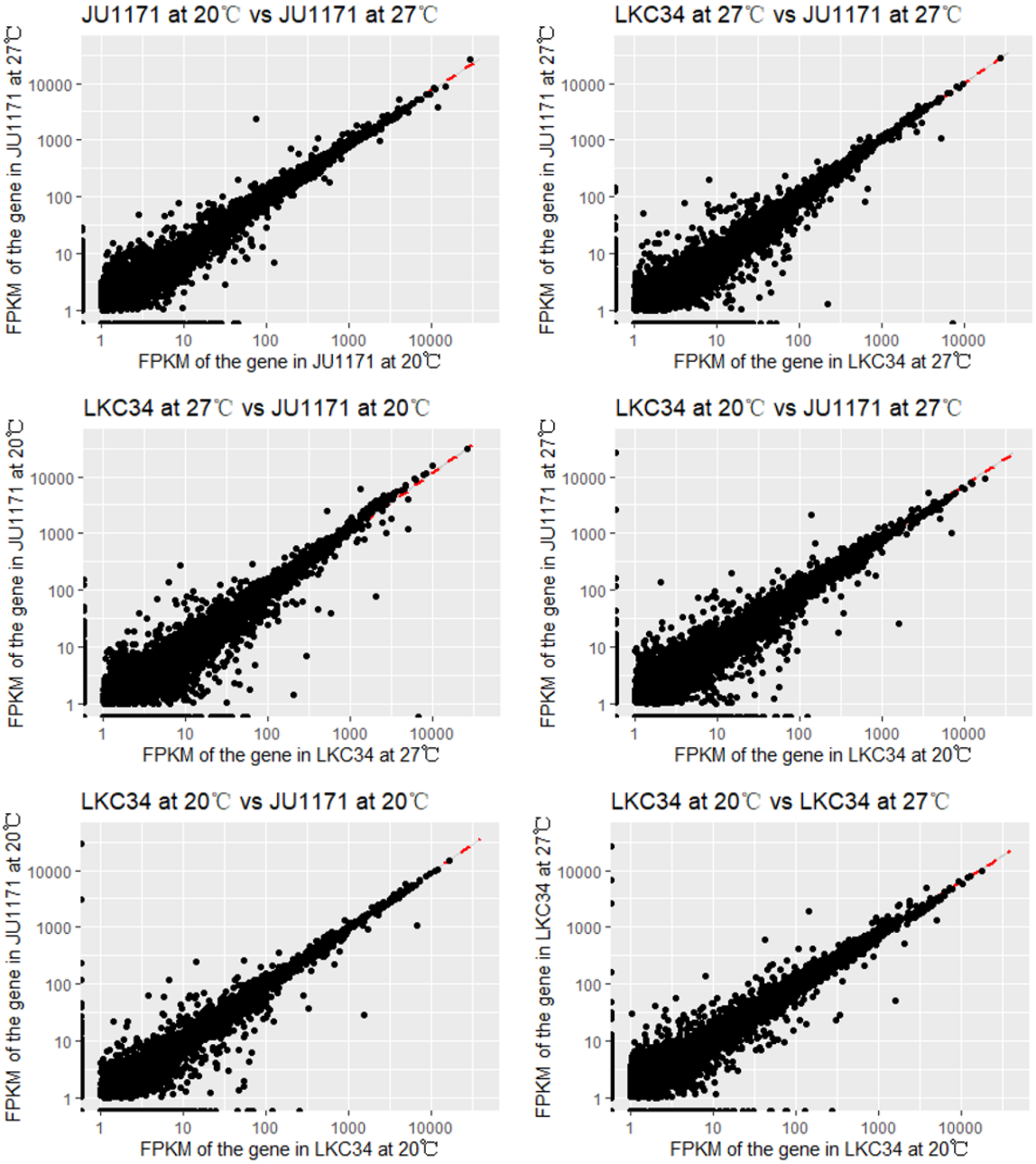


Figure 4. Scatter plots of differentially expressed genes in the six comparable groups. The red dashed lines are the regression lines. X-axis and Y-axis are the FPKM values of the gene in the two conditions respectively. Axes were rendered on the log10 scale.

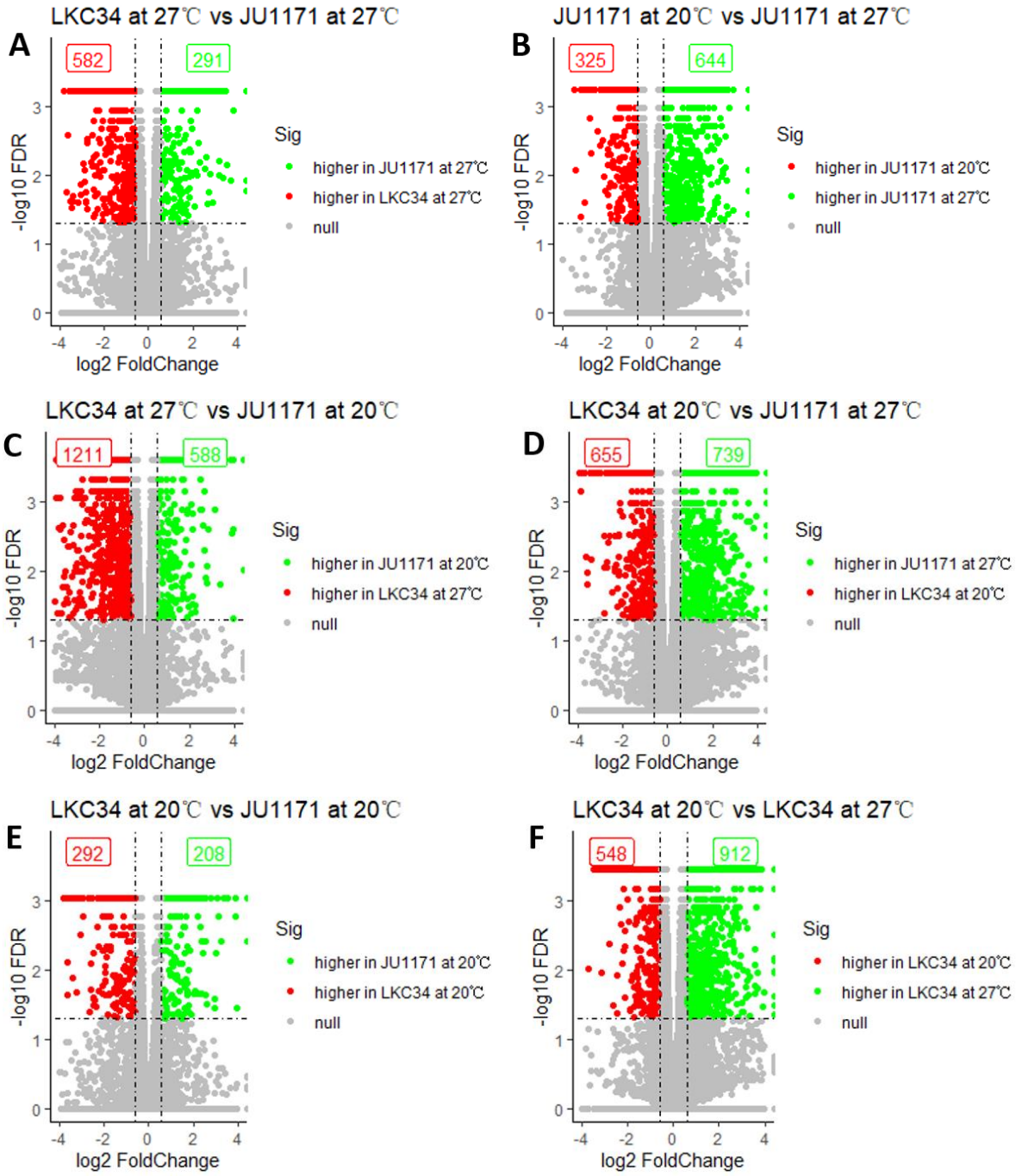


Figure 5. Volcano plots of differentially expressed genes in the six comparable groups generated by *Cuffdiff*. The horizontal lines represent the value of -log10 FDR where FDR = 0.05. The left vertical lines represent the value of log2 FoldChange where Fold Change = 1.5. The right vertical lines represent the value of log2 FoldChange where Fold Change = 1/1.5.

I manually made the six *gene_exp.diff* have the same gene IDs and in the same order and found that under the gene ID column, there were 869 cells that showed more

than one gene ID in one cell where it should just have one single gene ID in one cell. I did some research and found this case is due to the reason that these genes' positions are so close, they had some overlap so the *Cuffdiff* software cannot accurately detect which gene was represented by the reads. The *C.elegans* genome is very compact with often less than one kilobase between genes. In my thesis, I set a *q*-value cutoff of <0.05 for significant differential expression. After calling significantly differential expressed genes, there were still 616 cells that had more than one gene ID assigned it at least one of the six files. Because *Cuffdiff* had problems separating these genes, I moved on to other methods of differential gene expression analysis for my data.

3.5 Comparisons between *DESeq2*, *edgeR* and *limma*

Recently, the rapid output of high-throughput sequencing technologies for molecular genomic studies has led to an urgent need for statistical methods to quantify the differences between experiments for understanding the molecular basis of phenotype variation in biology. One of the most important aims is analyzing the RNA-Seq data to find the genes which are differentially expressed across multiple groups of samples between conditions. A number of statistical methods have been developed for RNA-Seq data based on Poisson and negative binomial distributions to detect the differential expressed genes (Park et al. 2016). According to Xiong et al. (2014), *edgeR*, *DESeq2*, *limma*-based methods, and *Cuffdiff* are among the most widely used tools for differential expression analysis. (NEW TABLE WOULD GO HERE – See note above)

Table 4. Comparisons between *DESeq2*, *edgeR* and *limma*

Method	Normalization	Read count distribution assumption	Differential expression test
DESeq2	DESeq sizeFactors	Negative Binomial distribution	Exact test
edgeR	trimmed mean of M values (TMM)	Negative Binomial distribution	Exact test
limma	trimmed mean of M values (TMM)	Voom transformation of counts	Empirical Bayes method

The first tool which was applied to our data is *DESeq2*. Love et.al (2014) presents *DESeq2*, which is a method for differential expression analysis of count data. It improved the stability and interpretability of estimates by using shrinkage estimators for dispersions and fold changes to enable more quantitative analysis. The *DESeq2* method detects and modifies low dispersion estimates by modeling the dependence of dispersion on average expression intensity in all samples. The package *DESeq2* provides methods to test for differential gene expression by using negative binomial generalized linear models (GLM) and uses local regression between mean and variance to estimate overdispersion. After GLMs were fitted for each gene, we used a Wald test in *DESeq2* for significant testing, where we used the estimated standard error of a log2 fold change to test if it was equal to zero. Besides, the likelihood ratio test (LRT) is also available as another option in *DESeq2*. With our data, we chose a cutoff of the adjusted *p*-value (*q*-value) of 0.05 and fold changes of 1.5 to call significantly differentially expressed gene lists. For visualization, six volcano plots were created to inspect differentially expressed genes generated by *DESeq2* (Figure 6)

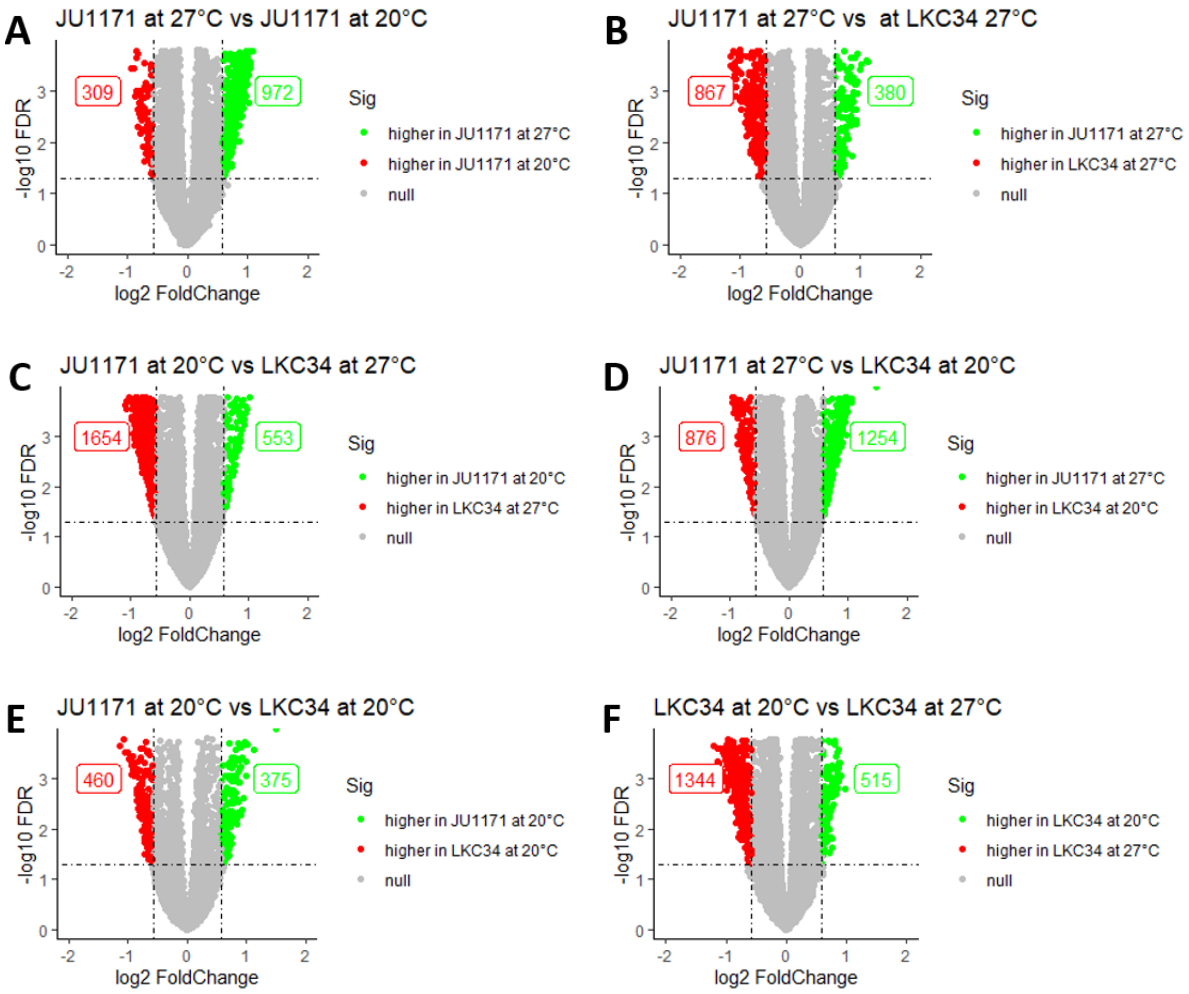


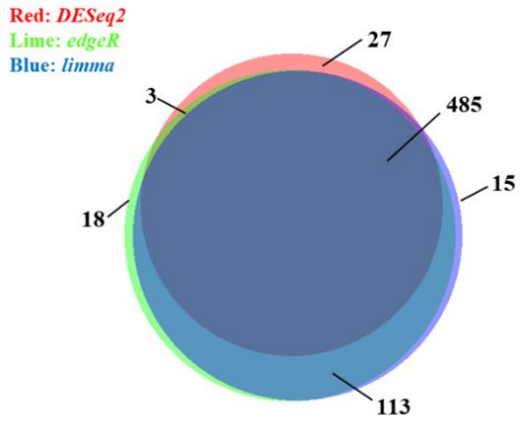
Figure 6. Volcano plots of differentially expressed genes in the six comparisons between groups generated by *DESeq2*. The horizontal lines represent the value of -log10 FDR where FDR = 0.05. The left vertical lines represent the value of log2 FoldChange where Fold Change = 1.5. The right vertical lines represent the value of log2 FoldChange where Fold Change = 1/1.5.

In addition, I used *edgeR* and compared it with the results with those of *DESeq2*. *edgeR* (empirical analysis of DGE in R) is a Bioconductor software package for examining the differential expression of replicated count data (Robinson et al. 2010). It is also based on the negative binomial generalized linear model and allows different sequencing depth by applying the Trimmed Mean of M values (TMM) normalization method. Empirical Bayes procedure was used to adjust for over-dispersion across genes (Park et al. 2016). For general experiments, once dispersion estimates were obtained and

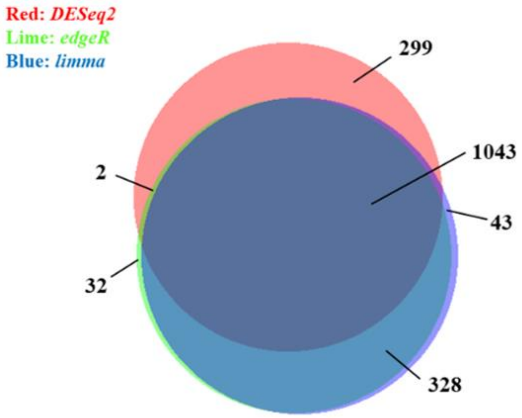
negative binomial generalized linear models were fitted, the differential expression was assessed for each gene using the quasi-likelihood (QL) F-test.

limma is also an R/Bioconductor software package that was originally designed for analyzing microarray data. To date, it has been extended to RNA-Seq data. *limma* is based on the gene-wise linear model and also uses TMM normalization for the adjustment of different sequencing depth. By using the empirical Bayes method, *limma* can deliver powerful inferences for differential expression analysis (Ritchie et al. 2015).

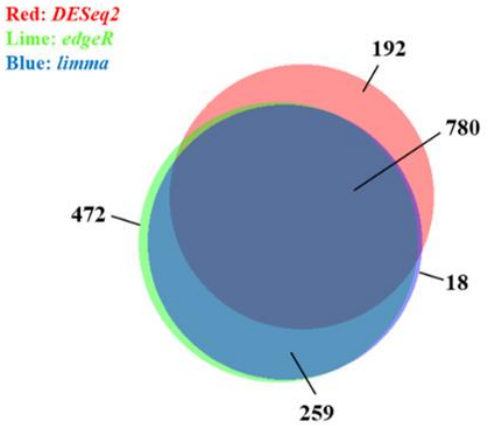
I applied *DESeq2, edgeR* and *limma* respectively, and then compared the results of these three methods which all use the same cutoff criteria of the adjusted *p*-value (*q*-value) of 0.05 and fold changes of 1.5 to call significantly differentially expressed gene lists. Venn-diagrams generated from BioVenn (http://www.biovenn.nl/) were used to display the overlap results between *DESeq2, edgeR* and *limma* (Figure 7 and Appendix 3). Take the genes that had higher expression in JU1171 at 20°C than in LKC34 at 20°C as an example, *DESeq2*, *edgeR* and *limma* identified 375, 412 and 418 genes as differentially expressed genes after FDR correction, separately. Among these, 304 genes were commonly detected in all of these three methods and 96 additional genes were commonly detected in both *edgeR* and *limma* methods (Appendix 3).



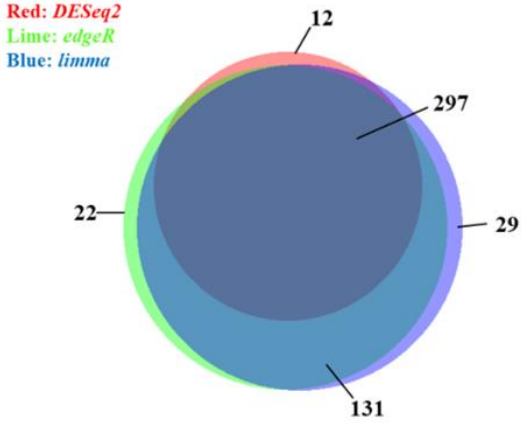
higher in LKC34 at 20°C than in LKC34 at 27°C



higher in LKC34 at 27°C than in LKC34 at 20°C



higher in JU1171 at 27°C than in JU1171 at 20°C



higher in JU1171 at 20°C than in JU1171 at 27°C

Figure 7. The Venn-diagrams were used to display the overlap among genes that were found to be significantly higher expressed under each condition by using the three differential expression testing tools: *DESeq2*, *edgeR* and *limma*. Only part of the diagram was shown, and the full part can be found in Appendix 3.

3.6 Comparisons between *DESeq2* and *DESeq*

In looking at the Volcano plot for our *DESeq2*, there are many genes that have a *q*-value of < 0.05 but do not have a significant fold change (Figure 6). This is a pattern that differs from some previously published RNA-Seq data from *C.elegans* dissected germlines (Campbell and Updike 2015). However, these other published data sets were analyzed using the *DESeq* pipeline instead of *DESeq2.* In order to determine if the large number of genes with a *q*-value of < 0.05 was due to the input data or the *DESeq2*

pipeline. I ran both a previously published data set (Campbell and Updike 2015) and my own data through both the *DESeq* and *DESeq2* pipelines.

Compared with *DESeq2*, which shrinks the gene-wise dispersion estimates towards the fitted values to obtain the final dispersion values, *DESeq* (Anders and Huber, 2010) adopts a more conservative approach using the maximum of the fitted value and the gene-wise estimate. The *DESeq* approach to test for the differential expression is very similar to the *edgeR* classic method, which uses an exact test for differences between two negative binomial variables. First, Campbell's raw mRNA-Seq data (Campbell and Updike 2015) were downloaded at the GEO database under the accession number GSE67954. The data from Campbell and Updike was from RNA isolated *C.elegans* germlines that were either wild type or depleted for the *csr-1* gene. I then ran the same mRNA-sequencing and analysis protocols as described in their paper, which included using *TopHat2* for alignment and *DESeq* for differential gene expression analysis. In addition, I used the mRNA-Seq analysis protocols as I used on my own data including *HISAT2* for alignment and *DESeq2* for differential gene expression analysis. Finally, I also ran *HISAT2* for alignment with *DESeq* for differential gene expression analysis. It can be seen from the volcano plots (Figure 8) that there was no big difference between *TopHat2* and *HISAT2*. The volcano plots for both pipelines using *DESeq* are similar to the published volcano plots with very few genes with an FDR of <0.05 that do not also have a >1.5 fold change in expression. However, the use of *DESeq2* resulted in an increase in the total number of up-regulated and down-regulated genes than *HISAT2* combined with *DESeq* (Figure 8). The use of *DESeq2* also resulted in a large number of genes with an FDR of <0.05 but not a >1.5 fold change.

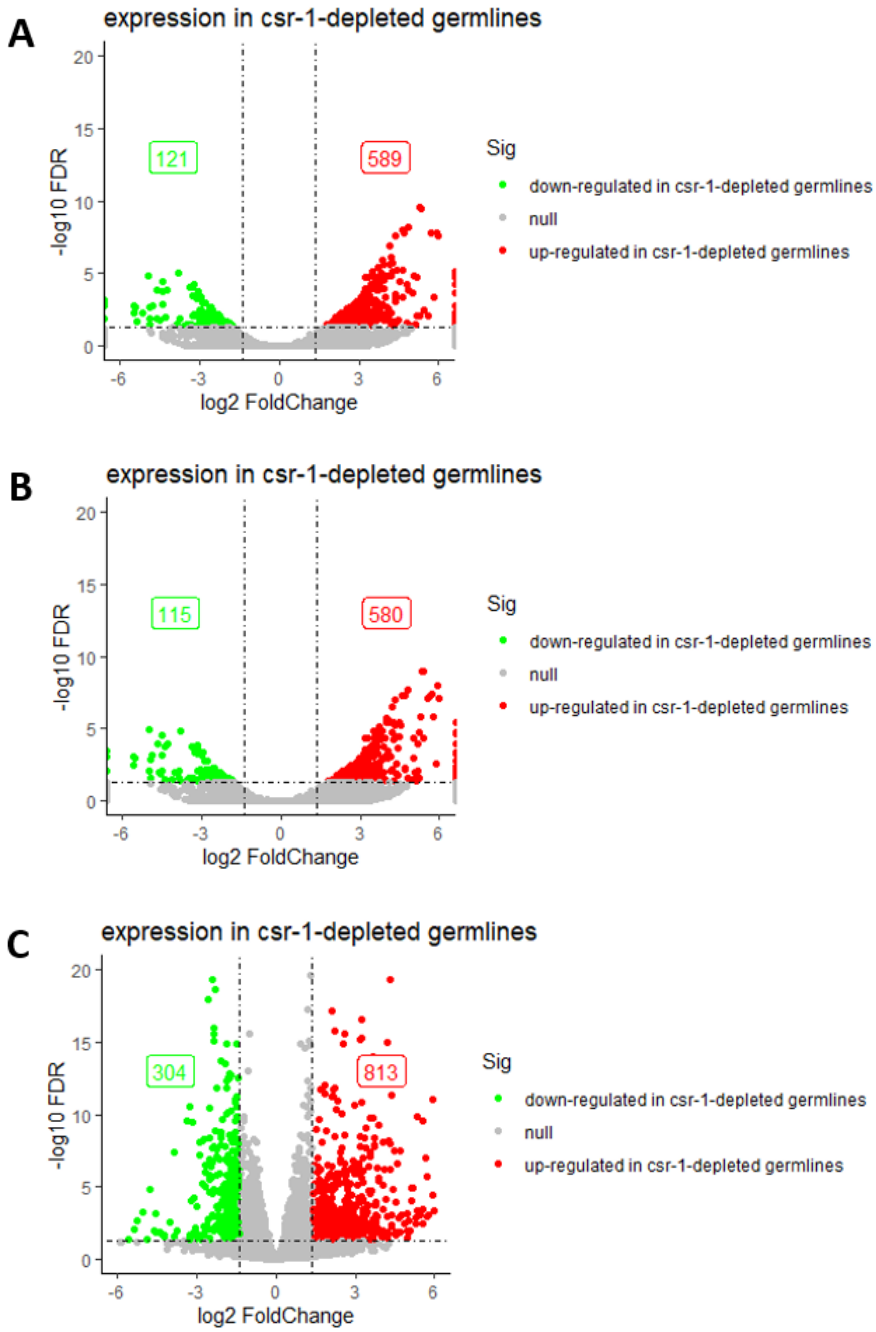


Figure 8. Volcano plots of differentially expressed genes with different protocols on partial of Campbell's mRNA-Seq data (Campbell and Updike 2015). (A). *TopHat2* + *htseq-count* + *DESeq*. (B). *Hisat2* + *htseq-count* + *DESeq*. (C). *Hisat2* + *htseq-count* + *DESeq2*. The horizontal lines represent the value of -log10 FDR where FDR = 0.05. The left vertical lines represent the value of log2 FoldChange where Fold Change = 1.5. The right vertical lines represent the value of log2 FoldChange where Fold Change = 1/1.5.

I next compared four differential gene expression pipelines on my own data:

DESeq, *DESeq2*, *edgeR* and *limma* using a single comparison of JU1171 versus LKC34

at 20°C. For all four of these pipelines, I used *HISAT2* for alignment. In general, *DESeq2*, *edgeR* and *limma* detected more differential expression genes than *DESeq*. Volcano plots were generated in R to visually present the differentially expressed genes (Figure 9). Venn-diagrams were used to display the overlap results between *DESeq* and *DESeq2* (Figure 10 and Appendix 4). Take the genes that are higher expressed in JU1171 at 20°C than in LKC34 at 20°C as an example, *DESeq* and *DESeq2* identified 268 and 375 genes as differentially expressed after FDR correction, respectively. Among these, 254 genes were commonly detected in both these two methods. While with all four methods, there was a high overlap of genes called as differentially expressed, the very different shapes of the volcano plots between methods points towards the differences in how the pipelines deal with determining both fold changes and FDR. Finally, all three newer methods (*DESeq2*, *edgeR,* and *limma)* all result in many more genes that have a <0.05 FDR without a significant fold change. These comparisons underscore the importance of taking the differential expression pipeline into account when comparing RNA-Seq experiments to previously published data which may have used a different pipeline.

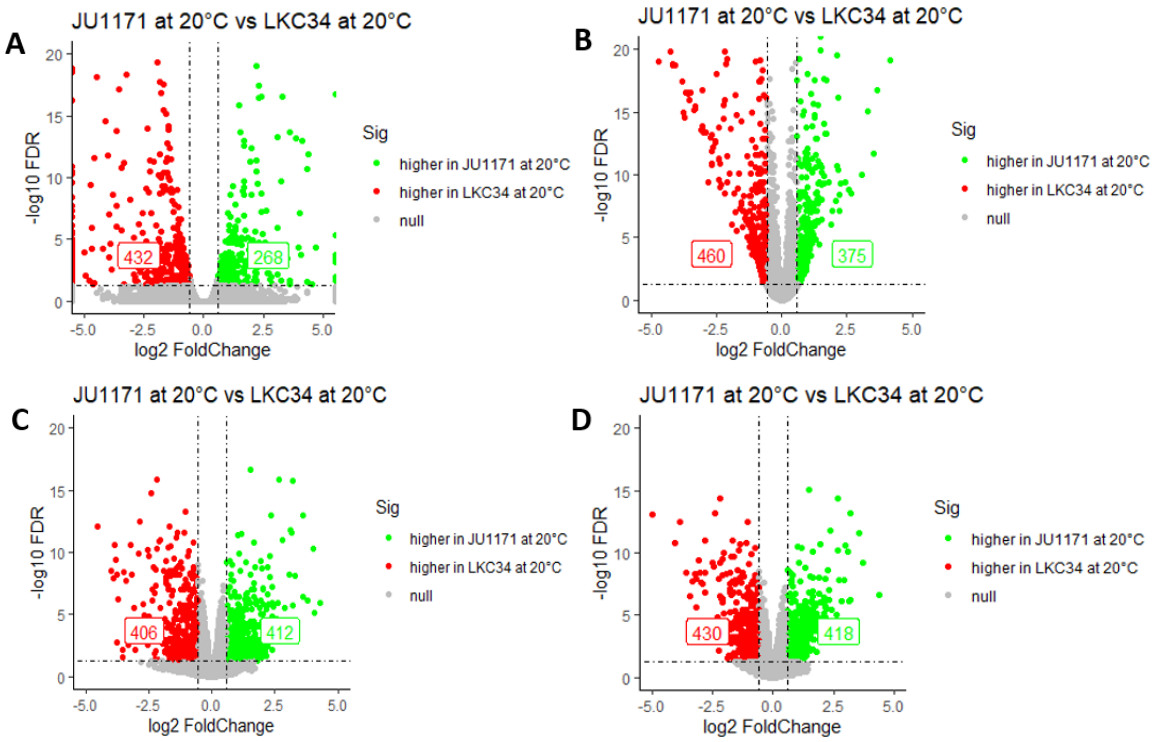


Figure 9. Volcano plots of differentially expressed genes with different protocols on partial of our data (JU20 versus LKC34 at 20°C). (A). *Hisat2 + htseq-count + DESeq*. (B). *Hisat2 + htseq-count + DESeq2*. (C). *Hisat2 + htseq-count + edgeR*. (D). *Hisat2 + htseq-count + limma*. The horizontal lines represent the value of -log10 FDR where FDR = 0.05. The left vertical lines represent the value of log2 FoldChange where Fold Change = 1.5. The right vertical lines represent the value of log2 FoldChange where Fold Change = 1/1.5.

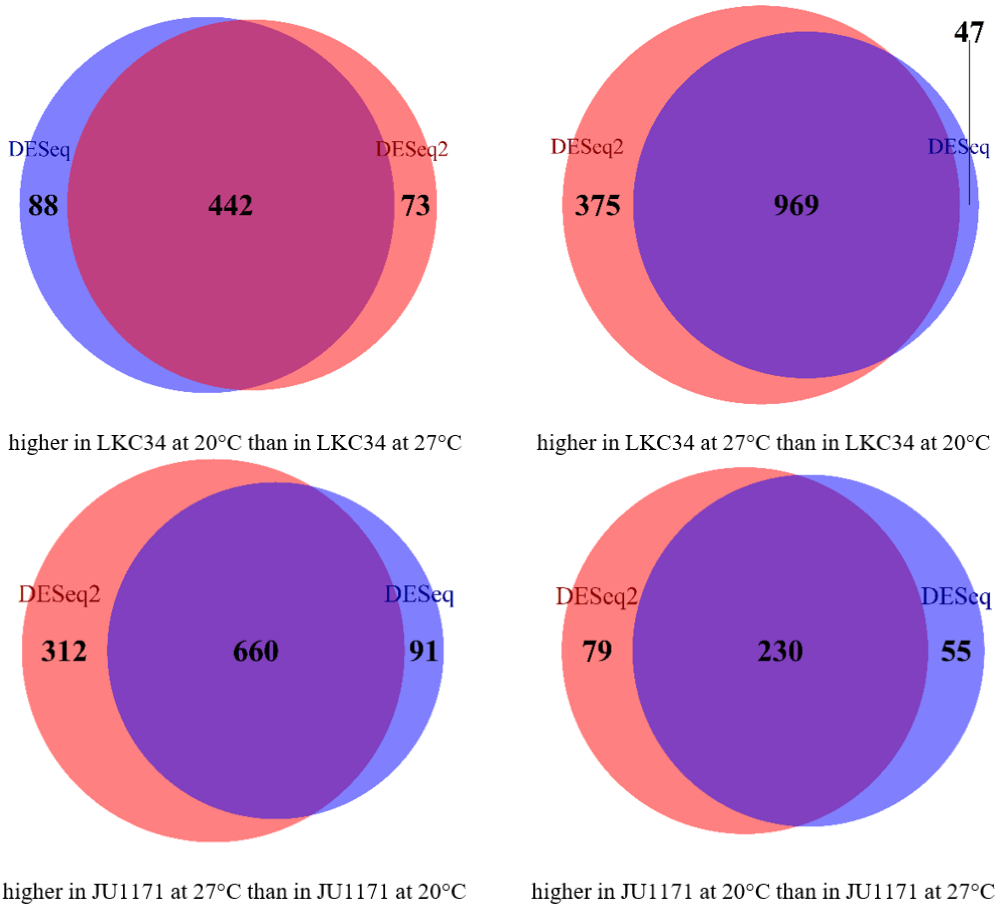


Figure 10. The Venn-diagrams were used to display the overlap among genes that were found to be significantly higher expressed under each condition by using the two differential expression testing tools: *DESeq* and *DESeq2*. Only part of the diagram was shown, and the full part can be found in Appendix 4.

3.7 Four Potential Patterns Analysis

The goal of analyzing these data sets was to find the molecular differences which could potentially underlie the reasons that JU1171 is more fertile than LK34 under the higher temperature condition. We propose that this difference in phenotype could be due to one of two scenarios: 1) there is a set of genes that change their expression with response to stress in JU1171 but not in LKC34, and/or 2) there is a set of genes with a baseline difference in expression between JU1171 and LKC34. Thus, I defined four different gene expression patterns in my thesis for these two scenarios (Figure 11).

Pattern 1 is those genes that are up-regulated in JU1171 at 27°C compared to 20°C but are not up-regulated in LKC34 at 27°C compared to 20°C (genes whose expressing in activated with elevated temperature in JU1171, but not in LKC34). Pattern 2 is those genes that are down-regulated in JU1171 at 27°C compared to 20°C but are not down-regulated in LKC34 at 27°C compared to 20°C (genes whose expression is down-regulated with elevated temperature in JU1171, but not in LKC34). Pattern 3 expressed higher in JU1171 compared to LKC34 at both temperatures and genes that are expressed higher in LKC34 compared to JU1171 at both temperatures (genes that always have a higher level of expression in JU1171 compared to LKC34). And Pattern 4 genes that are expressed higher in LKC34 compared to JU1171 at both temperatures (genes that always have a lower level of expression in JU1171 compared to LKC34). I determined the number of genes in each pattern by filtering the TURE of FALSE values for the corresponding columns in the results generated from *DESeq2* and got the gene ID lists for each pattern to do the following enrichment analysis. In total, there were 968 genes in Pattern 1, 305 genes in Pattern 2, 214 genes in Pattern 3 and 313 genes in Pattern 4.

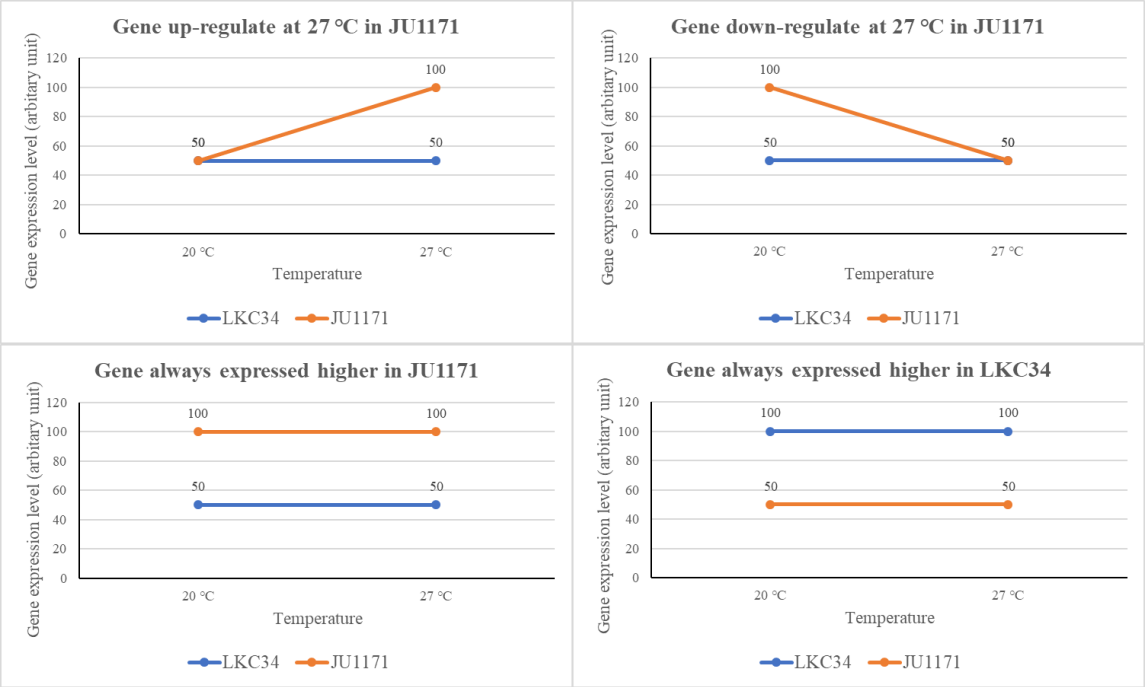


Figure 11. The line charts illustrate four different patterns based on different expression levels between LKC34 and JU1171 under different temperature conditions.

3.8 Gene Set Enrichment Analysis via the Hypergeometric Test

Gene set enrichment analysis is a method to find common characteristics of a set of genes of interests and it uses a statistical test to identify significantly enriched or depleted groups of genes (Subramanian et al., 2005). The hypergeometric test is a statistical test which uses the hypergeometric distribution to calculate the statistical significance to identify which sub-populations are over-represented or under-represented in a specific sample (Rivals et al. 2017). We wanted to test if there were any particular pathways or tissues for the genes based in our four patterns, which may be related to the potential reasons that JU1171 is more fertile at the higher temperature.

In my thesis research, I asked if there was enrichment for a specific tissue expression pattern in the four patterns by looking at the overlap of those genes with specific tissue expression pattern including genes whose expression is germline enriched

(Reinke et al. 2004), germline enriched gender neutral (Reinke et al. 2004), soma enriched (Reinke et al. 2004), neuron enriched (Watson et al. 2008), three different sets of spermatogenesis enriched genes (Reinke et al. 2004, Ortiz et al. 2014, Chu et al. 2006), and two sets of oocyte enriched genes (Reinke et al. 2004, Ortiz et al. 2014).

p-values for each gene set were calculated in R by applying the *phyper()* function and recorded in Table S1. Genes that had a *p*-value less than a given alpha are considered significant (Figure 12). Our default setting is an alpha of 0.01. Pattern 1 genes were significantly enriched for three gene sets, Spermatogenesis enriched genes (Reinke et al. 2004), Spermatogenesis enriched genes (Ortiz et al. 2014), and List of genes encoding spermatogenesis proteins (Chu et al. 2006) (Figure 12A). This can be interpreted that there were more genes among the 968 genes in Pattern 1 that were expressed in the sperm than expected. We also found that Pattern 1 genes were significantly under-represented in Germline enriched gender neutral (Reinke et al. 2004), Oocyte enriched genes (Reinke et al. 2004) and Oocyte enriched genes (Ortiz et al. 2014). This can be explained that there were fewer genes among the 968 genes in Pattern 1 that were expressed in oocyte than expected. This shift towards a spermatogenic pattern of expression away from an oogenic pattern of expression may reflect a difference in the stages of germline development in JU1171 versus LKC34. Under normal circumstances, the L4 stage of *C. elegans* makes sperm, while adult *C. elegans* make oocytes. Perhaps this transition occurs latter in JU1171, which allows them to make more sperm and be more fertile.

For Pattern 2, genes were significantly enriched for Oocyte enriched genes (Ortiz et al. 2014), which can be interpreted that there were more genes among the 305 genes in Pattern 2 that were expressed in the oocyte than expected (Figure 12B). In addition,

Pattern 2 genes were significantly under-represented in Germline enriched genes (Reinke et al. 2004), Spermatogenesis enriched genes (Reinke et al. 2004), Spermatogenesis enriched genes (Ortiz et al., 2014) and Oocyte enriched genes (Reinke et al. 2004). It can be explained that there were fewer genes among the 305 genes in Pattern 2 that were expressed in germline and sperm than expected.

In addition, Patter 4 genes were significantly enriched in Germline enriched gender neutral gene set (Reinke et al. 2004) and Oocyte enriched gene set (Ortiz et al. 2014). It can be interpreted that there were fewer genes among the 313 genes in Pattern 4 that were expressed in germlines and oocyte than expected. Interestingly, we got no significant *p*-value in Pattern 3 among all the public gene sets. The full table can be found in Appendix 5.

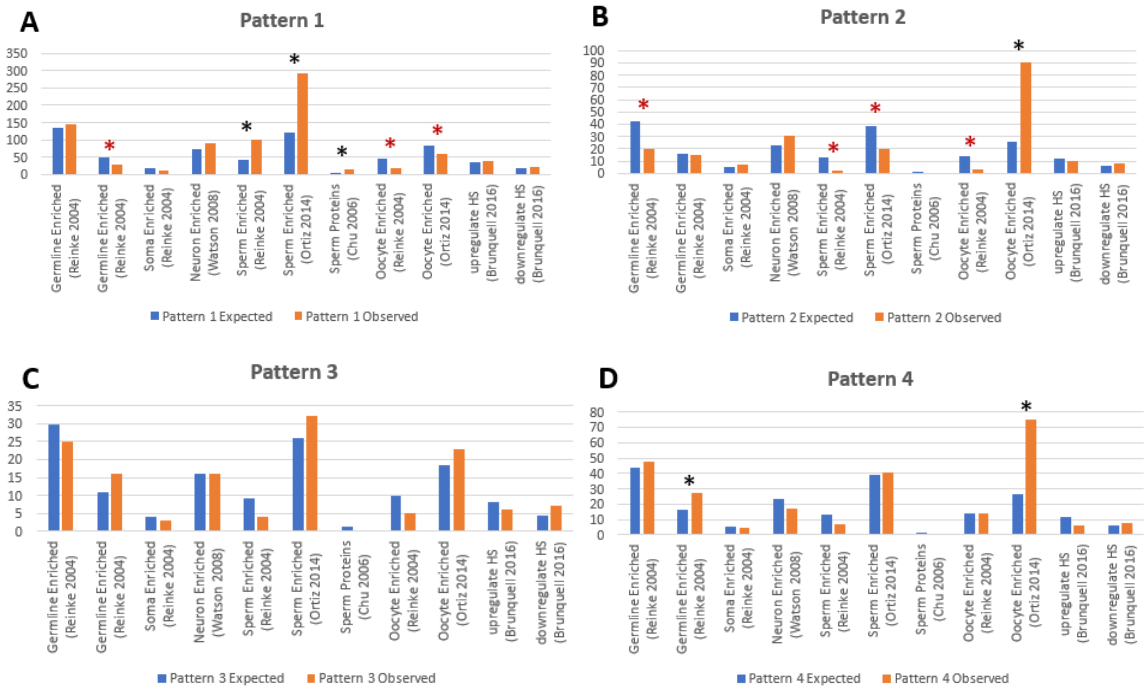


Figure 12. The bar charts illustrate the hypergeometric test results of four potential patterns. The X-axis is the abbreviation of eleven published gene sets' names. From left to right: germline enriched genes (Reinke et al. 2004), Germline enriched gender neutral (Reinke et al. 2004), Soma enriched genes (Reinke et al. 2004), Neuron enriched genes (Watson et al. 2008), Spermatogenesis enriched genes (Reinke et al. 2004), Spermatogenesis enriched genes (Ortiz et al. 2014), List of genes encoding spermatogenesis proteins (Chu et al. 2006), Oocyte enriched genes from (Reinke et al. 2004), Oocyte enriched genes from (Ortiz et al. 2014), List of the significantly up-regulated genes altered in response to the "hsf-1(+);+HS vs control" condition (Brunquell et al. 2016) and list of the significantly down-regulated genes altered in response to the "hsf-1(+);+HS vs control" condition (Brunquell et al. 2016). Y-axis is the number of expected and observed genes of each pattern. *$P < 0.01$. The black asterisk represents the genes that were significantly enriched in each dataset, while the red asterisk represents the genes that were significantly under-represented in each dataset.

We now know that organisms have evolved an ancient heat shock response (HSR) to protect cells at elevated temperatures (Schreine et.al 2019). This response was driven by the heat shock transcription factor (HSF1). The HSF1 homolog HSF-1 in *C. elegans* is an important protein that is required to activate a stress-dependent response (Brunquell et.al 2016). We wanted to test if there was enrichment for HSF-1 responsive genes in the four patterns, which would indicate that in JU1171 HSR is activated so that it can be more fertile at the higher temperature. We looked at the overlap of the genes in the four

patterns with a list of the significantly up-regulated genes altered in response to the “hsf-1(+);+Heat-Shock vs control” condition (Brunquell et al. 2016) and a list of the significantly down-regulated genes altered in response to the “hsf-1(+);+Heat-Shock vs control” condition (Brunquell et al. 2016). There is no significant *p*-value in all of the four defined patterns among these two gene sets.

3.9 Gene Ontology (GO) Analysis

One of the main uses of the GO is to perform enrichment analysis on gene sets (Ashburner et al. 2000). For instance, given four sets of genes based on our four patterns, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for those gene sets. The gene ID lists were used to do the enrichment analysis on the functions of genes under each pattern to uncover the expression patterns that are different between the two strains and two temperatures deeply. The Gene Ontology (GO) knowledgebase is the world’s largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research. The results page of using the PANTHER Overrepresentation Test displayed a table that lists significant enriched GO terms (or parents of GO terms) used to describe the set of genes that we entered on the previous page.

Further analysis using PANTHER Overrepresentation Test resulted in 46 overrepresented Gene Ontology (GO) biological process terms (Table S2) for Pattern 1. The five most enriched GO terms were neuron migration (GO:0001764), positive

regulation of neurogenesis (GO:0050769), positive regulation of nervous system development (GO:0051962), neuron projection development (GO:0031175), and molting cycle (GO:0042303). There are no statistically significant results for the other three patterns.

4. DISCUSSION

4.1 Read Alignment Rates and Processing Time Indicate that *HISAT2* Works Better than *TopHat2*

My study did several comparisons between *TopHat2* and *HISAT2* with different parameter settings to come to an optimal one. I tested 10 groups for *HISAT2* and 19 groups for *TopHat2* and selected the most optimal one based on both the alignment rate and the running time. However, there were two biases in my test. First, I only ran these comparisons using a single test file and not on the full 16 datasets, in order to choose the one I eventually chose was the best. But this parameter setting may not be the optimal one and may not be applicable for other datasets. Second, *HISAT2* has a unique option, "*--dta-cufflinks*", which can report alignments tailored specifically for *Cufflinks.* I did add this option when I run the *Cuffdiff* pipeline. However, when I first started to look for the best parameters, I did not add this option. Thus, there could be some slight differences in the alignment rates between the test results and the final results of the *HISAT2* used for *Cuffdiff* protocol. We gave up using *TopHat2* not only because of its lower alignment rates and long runtimes compared to the *HISAT2*, but also because it has entered a low maintenance, low support stage as it is now superseded by *HISAT2* more accurately and efficiently.

4.2 No Method among *Cuffdiff*, *DESeq2*, *edgeR*, *limma* and *DESeq* is Optimal under All Circumstances

In this master thesis, I compared several methods for calling differential gene expression analysis based on RNA-Seq data. I applied *Cuffdiff* and also the most widely

used methods that are available in R or Bioconductor, which are *DESeq2*, *edgeR*, *limma* and *DESeq*. My key purpose was to come to a sound recommendation on which methods performed better than others and could give us the most conservative result for our 16 samples. In my thesis, just a brief summary of these software packages that I compared was given. For a more detailed description of the packages and the introduction of the statistical methods they apply, people can refer to their original publication works of literature and websites. When applying these methods, I followed the instructions and used the recommended approach that an average user is likely to use, which includes the common parameters and default normalization methods (Seyednasrollah et.al 2015).

I did more theory comparisons among those methods but not more parameter settings or detailed results comparisons, such as run times and the effect of normalization on the detections. I compared their final results by seeing their overlaps using Venn-diagrams and most of the figures showed that *edgeR* and *limma* gave us the most similar results, while with *DESeq2* there were more uniquely called differentially expressed genes. It might because I took the recommendation from the *limma* user guide to use TMM normalization of the *edgeR* package. *DESeq2* internally corrects for library size by estimating the size factors for each sample using *DESeq()* function. The reason why I finally chose the results from *DESeq2* is not only because most of the literature about RNA-Seq analysis for *C.elegans* used *DESeq* or *DESeq2* as the main tool, such as Campbell's paper (Campbell and Updike 2015), but also because *DESeq2* has been developed to deal with the analysis of experiments with a small number of replicates and even work with experiments without any biological replicates (Seyednasrollah et.al 2015). Moreover, *DESeq2* enables the shrinkage of effect size by introducing *lfcShrink*()

function, which performs log2 fold change shrinkage and is useful for visualization and ranking of genes. Overall, from what I have seen in my own testing, the *DESeq2, edgeR, limma* and *DESeq* typically report a high percentage of overlapping sets of differentially expressed genes and have similar performance for the differential expression testing on our 16 samples.

4.3 Gaps in How the Genes were Changed at Elevated Temperature

The temperature has long been thought to regulate lifespan by globally affecting biological processes and chemical reactions. Generally, it is believed that lower temperatures prolong lifespan while higher temperatures shorten it. However, recent work demonstrated that germline function is more buffered in JU1171 and LKC34 at higher temperature conditions (Petrella 2014). Our studies have uncovered some genes that may play an active role in temperature modulation of *C.elegans* germline function. GO analysis gave us the results that several particular related pathways and tissues in our four patterns of genes that may help JU1171 survive at the higher temperature, such as the regulation of neurogenesis, nervous system development and molting cycle.

With pattern 1 analysis, there was both an over-enrichment of genes associated with spermatogenesis (tissue expression comparison results) and an over-enrichment of genes associated with neurons (GO analysis results). Both of these two categories are the categories that are over-enriched with genes that are up-regulated in germlines when P-granules are lost (Knutson et al. 2017; Campbell and Updike 2015). Therefore, there may be a link between decreased P-granules and the better fertility in JU1171 at higher temperatures. Interestingly, there is increased germline apoptosis when P-granules are

lost (Min et al. 2016). Previous research has shown that there is an increase in germline apoptosis at 26.5°C (Poullet et al. 2015). Based on those findings, we predicted that the strains that are better able to increase germline apoptosis may be able to protect oocyte quality, thus leading to increased fertility. If there is a bigger issue with P-granules in JU1171, it may lead to more apoptosis which might explain why JU1171 is more fertile under the higher temperature condition.

We now know that all of the biological processes mentioned above may assist JU1171 survive at the higher temperature. But we still do not know exactly how these genes were changed in those biological processes. For future work, more wet-lab experiments should be done to investigate these differentially expressed genes and focus more on the relevant pathways when we see that in other analyses. Involving more gene expression results of extra worms may also significantly improve the accuracy of our analysis.

5. BIBLIOGRAPHY

Allaire, J. "RStudio: integrated development environment for R." *Boston, MA* 537 (2012): 538.

Amrit, Francis RG, and Arjumand Ghazi. "Transcriptomic Analysis of C. elegans RNA Sequencing Data Through the Tuxedo Suite on the Galaxy Project." *JoVE (Journal of Visualized Experiments)* 122 (2017): e55473.

Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." *Nature Precedings* (2010): 1-1.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data." *Bioinformatics* 31.2 (2015): 166-169.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995): 289-300.

Brenner, Sydney. "The genetics of Caenorhabditis elegans." *Genetics* 77.1 (1974): 71-94.

Brunquell, Jessica, et al. "The genome-wide role of HSF-1 in the regulation of gene expression in Caenorhabditis elegans." *BMC genomics* 17.1 (2016): 559.

Campbell, Anne C., and Dustin L. Updike. "CSR-1 and P granules suppress sperm-specific transcription in the C. elegans germline." *Development* 142.10 (2015): 1745-1755.

Chen, Yunshun, Aaron TL Lun, and Gordon K. Smyth. "Differential expression analysis of complex RNA-seq experiments using edgeR." *Statistical analysis of next generation sequencing data*. Springer, Cham, 2014. 51-74.

Conesa, Ana, et al. "A survey of best practices for RNA-seq data analysis." *Genome biology* 17.1 (2016): 13.

Gómez-Orte, Eva, et al. "Effect of the diet type and temperature on the C. elegans transcriptome." *Oncotarget* 9.11 (2018): 9556.

Harvey, S. C., and M. E. Viney. "Thermal variation reveals natural variation between isolates of Caenorhabditis elegans." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 308.4 (2007): 409-416.

Ji, Fei, and Ruslan I. Sadreyev. "RNA-seq: Basic bioinformatics analysis." *Current protocols in molecular biology* 124.1 (2018): e68.

Knutson, Andrew Kekūpa'A., et al. "Germ granules prevent accumulation of somatic transcripts in the adult Caenorhabditis elegans germline." *Genetics* 206.1 (2017): 163-178.

Law, Charity W., et al. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." *Genome biology* 15.2 (2014): R29.

Leung, Maxwell CK, et al. "Caenorhabditis elegans: an emerging model in biomedical and environmental toxicology." *Toxicological sciences* 106.1 (2008): 5-28.

Liao, Yang, Gordon K. Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." *Bioinformatics* 30.7 (2014): 923-930.

Liao, Yang, Gordon K. Smyth, and Wei Shi. "The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads." *Nucleic acids research* 47.8 (2019): e47-e47.

Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome biology* 15.12 (2014): 550.

Lun, Aaron TL, Yunshun Chen, and Gordon K. Smyth. "It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR." *Statistical Genomics*. Humana Press, New York, NY, 2016. 391-416.

McDermaid, Adam, et al. "Interpretation of differential gene expression results of RNA-seq data: review and integration." *Briefings in bioinformatics* 20.6 (2019): 2044-2054.

Min, Hyemin, Yhong-Hee Shim, and Ichiro Kawasaki. "Loss of PGL-1 and PGL-3, members of a family of constitutive germ-granule components, promotes germline apoptosis in C. elegans." *J Cell Sci* 129.2 (2016): 341-353.

Ortiz, Marco A., et al. "A new dataset of spermatogenic vs. oogenic transcriptomes in the

nematode Caenorhabditis elegans." *G3: Genes, Genomes, Genetics* 4.9 (2014): 1765-1772.

Park, Hyunjin, et al. "Multivariate approach to the analysis of correlated RNA-seq data." *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016.

Pertea, Mihaela, et al. "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown." *Nature protocols* 11.9 (2016): 1650.

Petrella, Lisa N. "Natural variants of C. elegans demonstrate defects in both sperm function and oogenesis at elevated temperatures." *PLoS One* 9.11 (2014). Brenner S, 1974 Conesa Genetics 77: 71–94.

Phipson, Belinda, et al. "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression." *The annals of applied statistics* 10.2 (2016): 946.

Poullet, Nausicaa, et al. "Evolutionarily divergent thermal sensitivity of germline development and fertility in hermaphroditic Caenorhabditis nematodes." *Evolution & development* 17.6 (2015): 380-397.

Prasad, Anisha, et al. "Temperature-dependent fecundity associates with latitude in Caenorhabditis briggsae." *Evolution: International Journal of Organic Evolution* 65.1 (2011): 52-63.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rechtsteiner, Andreas, et al. "Repression of germline genes in Caenorhabditis elegans somatic tissues by H3K9 Dimethylation of their promoters." *Genetics* 212.1 (2019): 125-140.

Ritchie, Matthew E., et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic acids research* 43.7 (2015): e47-e47.

Rivals, Isabelle, et al. "Enrichment or depletion of a GO category within a class of genes: which test?." *Bioinformatics* 23.4 (2007): 401-407.

Robinson, Mark D., and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data." *Genome biology* 11.3 (2010): R25.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26.1 (2010): 139-140.

Schreiner, William P., et al. "Remodeling of the Caenorhabditis elegans non-coding RNA transcriptome by heat shock." *Nucleic acids research* 47.18 (2019): 9829-9841.

Seyednasrollah, Fatemeh, Asta Laiho, and Laura L. Elo. "Comparison of software packages for detecting differential expression in RNA-seq studies." *Briefings in bioinformatics* 16.1 (2015): 59-70.

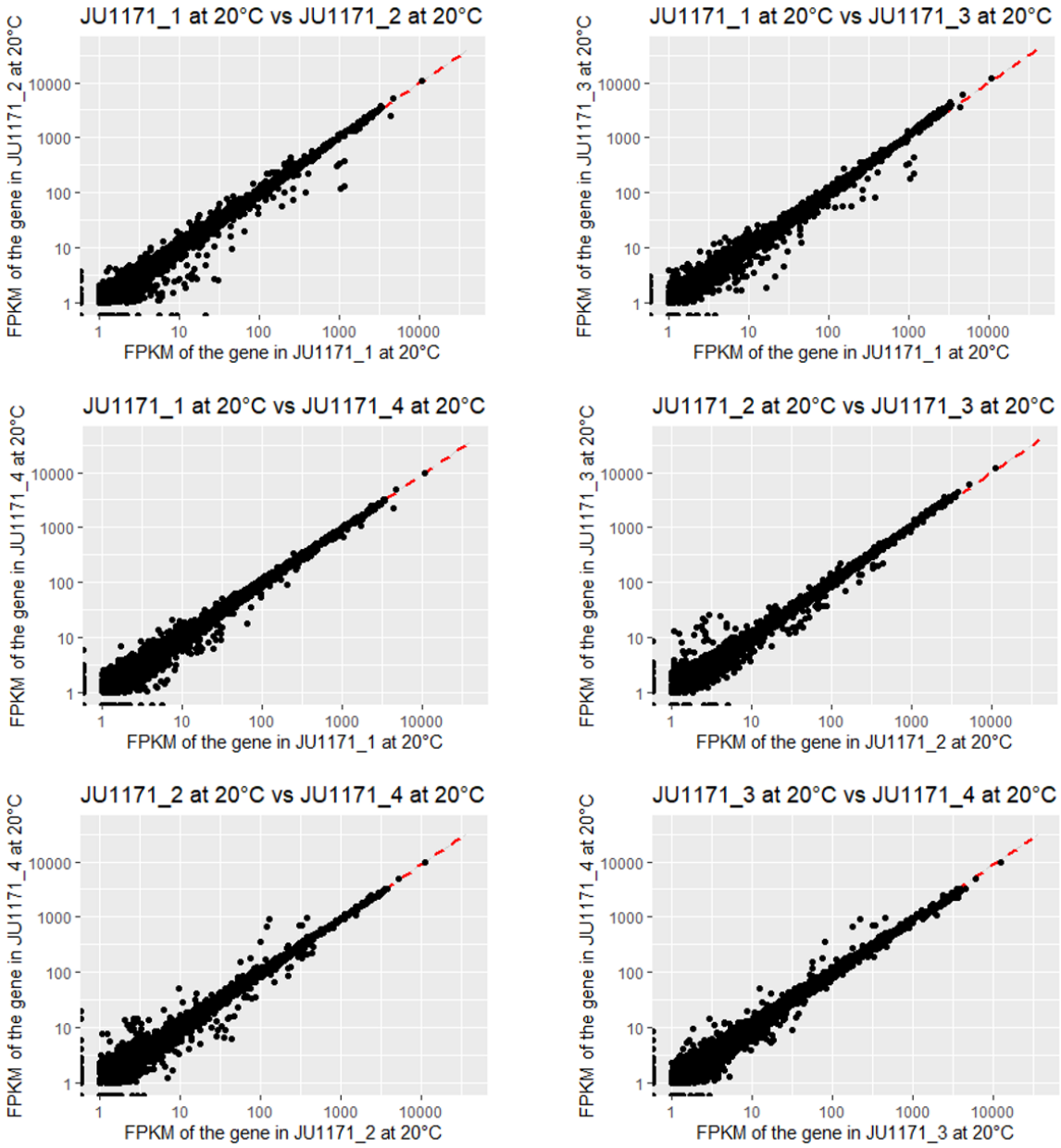
Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences* 102.43 (2005): 15545-15550.

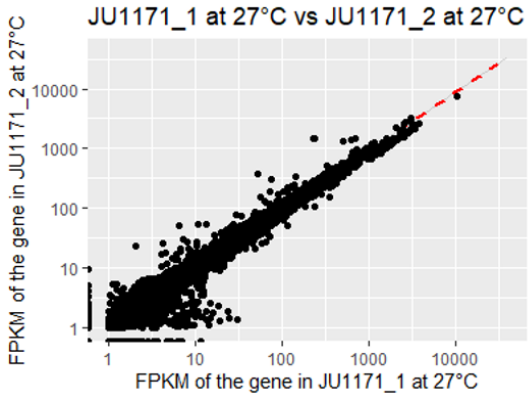
Trapnell, Cole, et al. "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." *Nature protocols* 7.3 (2012): 562-578.

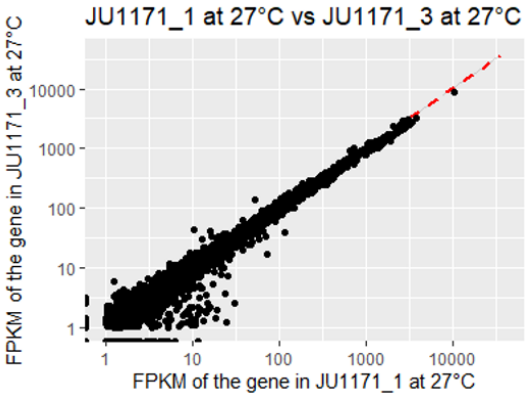
Xiong, Hao, et al. "DE-FPCA: testing gene differential expression and exon usage through functional principal component analysis." *Statistical analysis of next generation sequencing data*. Springer, Cham, 2014. 129-143.

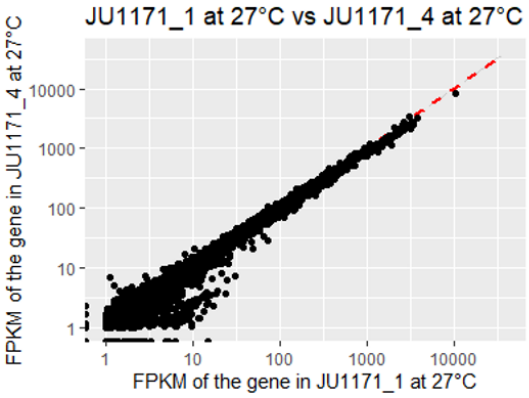
6. APPENDIX

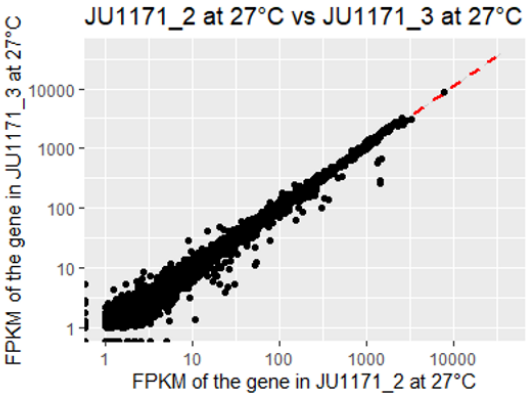
Appendix 1.

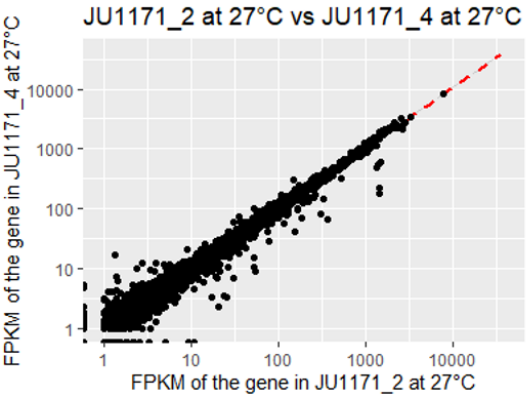


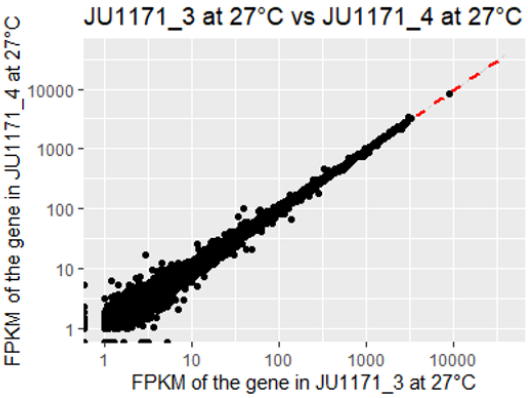

JU1171_1 at 27°C vs JU1171_2 at 27°C
FPKM of the gene in JU1171_2 at 27°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

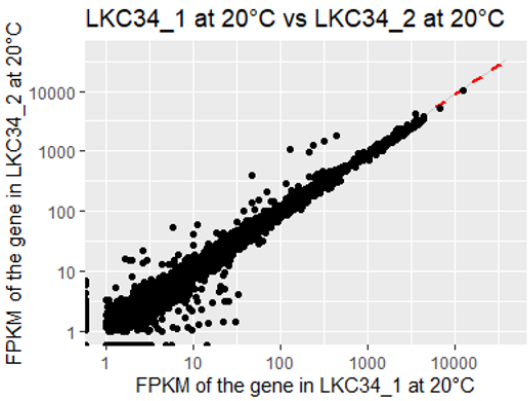

JU1171_1 at 27°C vs JU1171_3 at 27°C
FPKM of the gene in JU1171_3 at 27°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

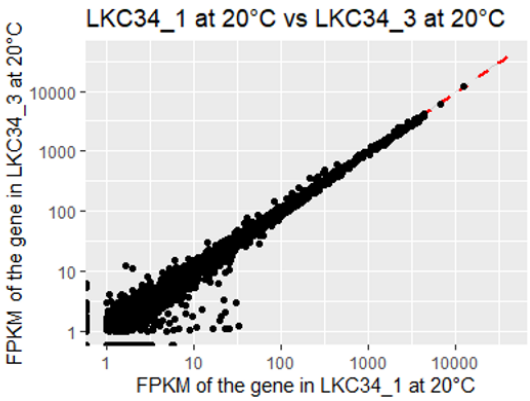

JU1171_1 at 27°C vs JU1171_4 at 27°C
FPKM of the gene in JU1171_4 at 27°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

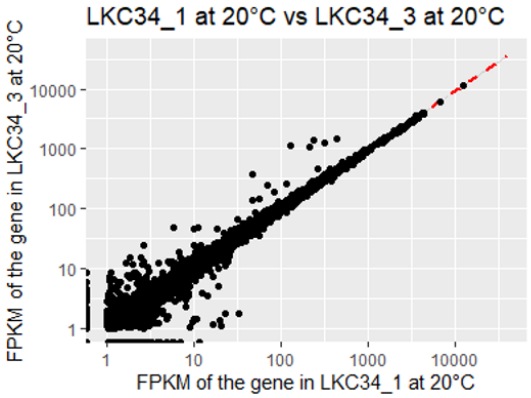

JU1171_2 at 27°C vs JU1171_3 at 27°C
FPKM of the gene in JU1171_3 at 27°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

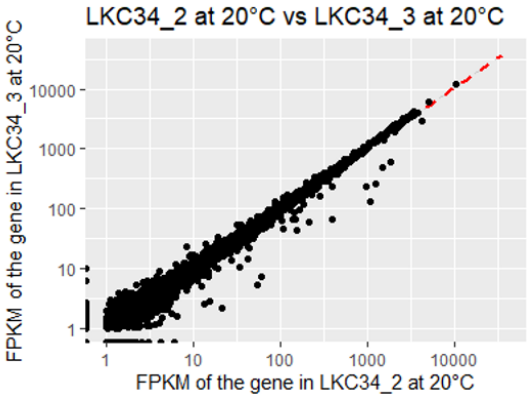

JU1171_2 at 27°C vs JU1171_4 at 27°C
FPKM of the gene in JU1171_4 at 27°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

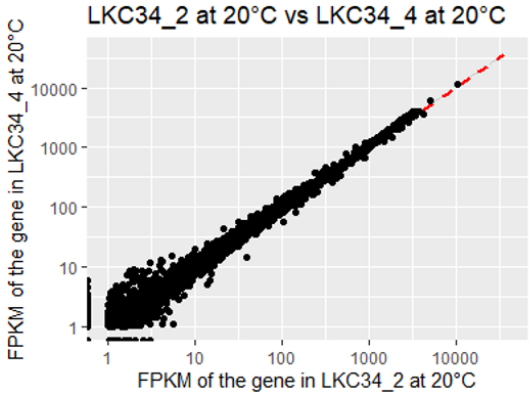

JU1171_3 at 27°C vs JU1171_4 at 27°C
FPKM of the gene in JU1171_4 at 27°C
FPKM of the gene in JU1171_3 at 27°C
1
10
100
1000
10000

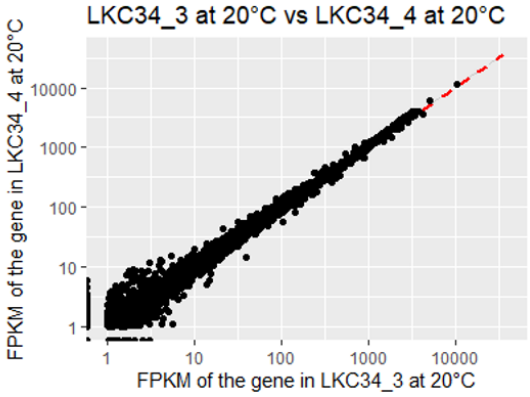

LKC34_1 at 20°C vs LKC34_2 at 20°C
FPKM of the gene in LKC34_2 at 20°C
FPKM of the gene in LKC34_1 at 20°C
1
10
100
1000
10000

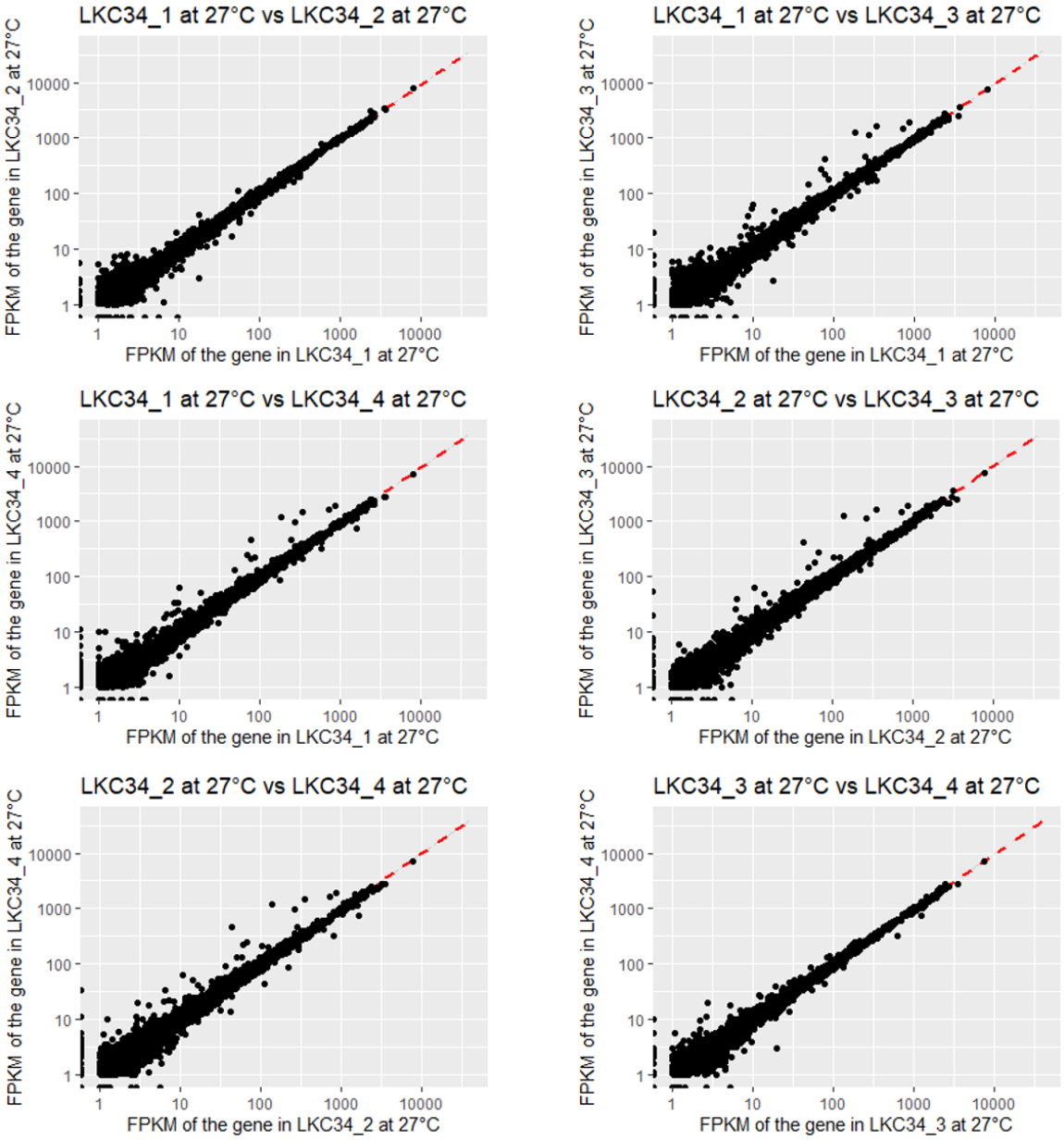

LKC34_1 at 20°C vs LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 20°C
FPKM of the gene in LKC34_1 at 20°C
1
10
100
1000
10000


LKC34_1 at 20°C vs LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 20°C
FPKM of the gene in LKC34_1 at 20°C
1
10
100
1000
10000


LKC34_2 at 20°C vs LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 20°C
FPKM of the gene in LKC34_2 at 20°C
1
10
100
1000
10000

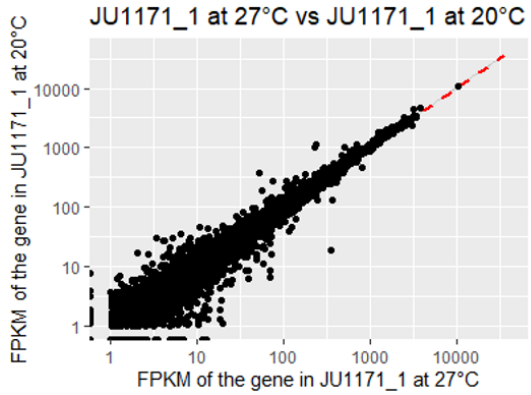

LKC34_2 at 20°C vs LKC34_4 at 20°C
FPKM of the gene in LKC34_4 at 20°C
FPKM of the gene in LKC34_2 at 20°C
1
10
100
1000
10000

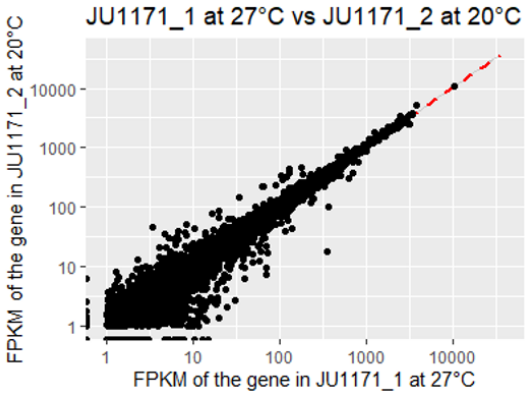

LKC34_3 at 20°C vs LKC34_4 at 20°C
FPKM of the gene in LKC34_4 at 20°C
FPKM of the gene in LKC34_3 at 20°C
1
10
100
1000
10000

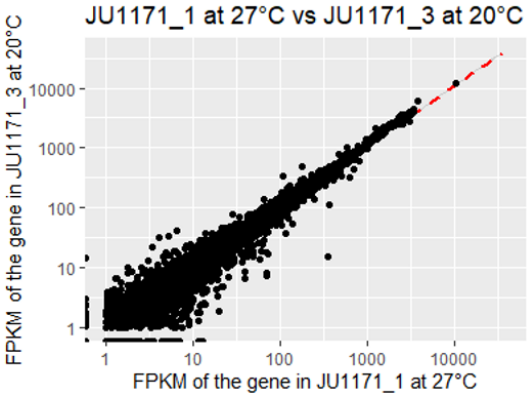


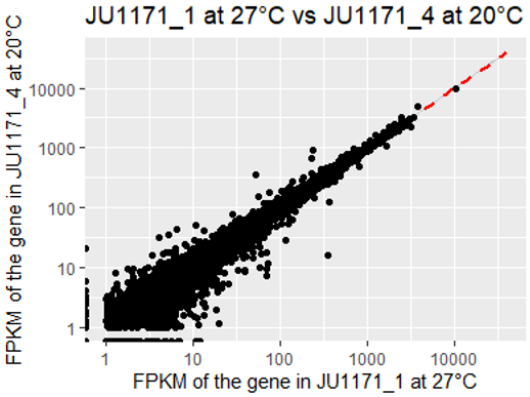
Scatter plots were created to compare the gene expression of each replicate for JU1171 at 20°C, JU1171 at 27°C, LKC34 at 20°C and LKC34 at 27°C. The red dashed lines are the regression lines. X-axis and Y-axis are the FPKM values of the gene in the two replicates respectively. Axes were rendered on the log10 scale.

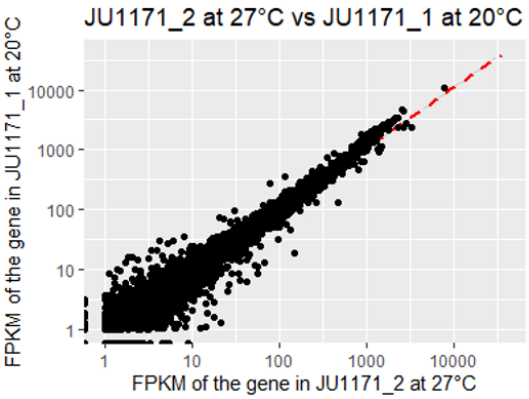
Appendix 2.

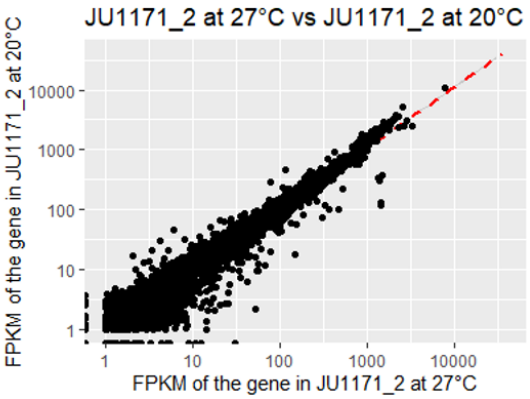

JU1171_1 at 27°C vs JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

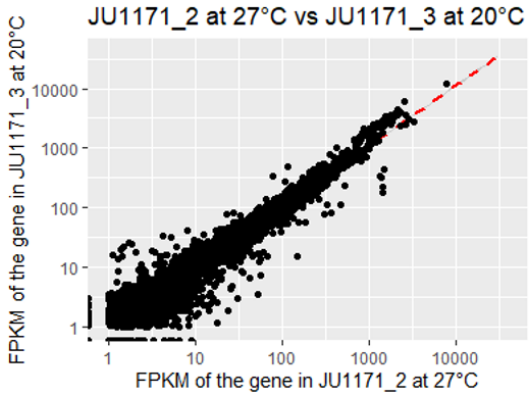

JU1171_1 at 27°C vs JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

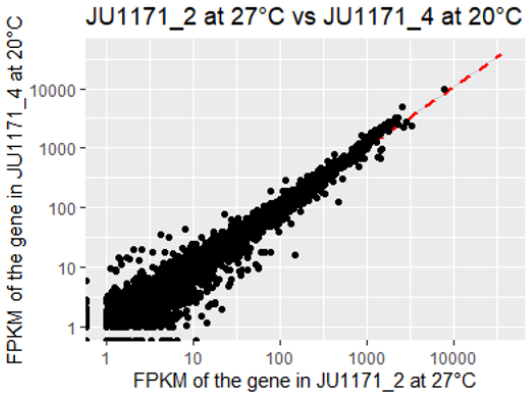

JU1171_1 at 27°C vs JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

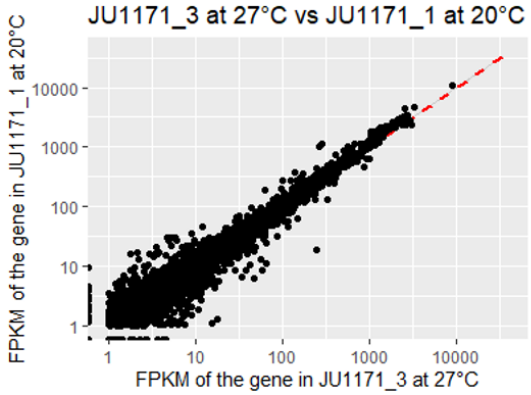

JU1171_1 at 27°C vs JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 20°C
FPKM of the gene in JU1171_1 at 27°C
1
10
100
1000
10000

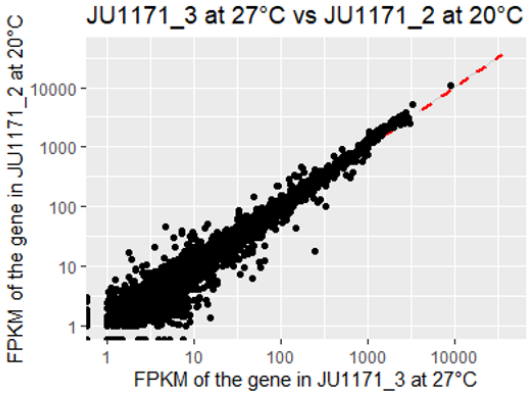

JU1171_2 at 27°C vs JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

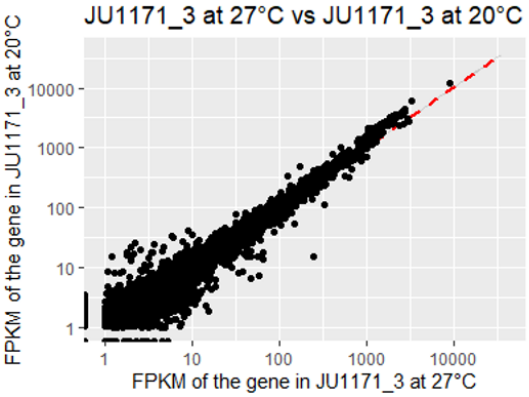

JU1171_2 at 27°C vs JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

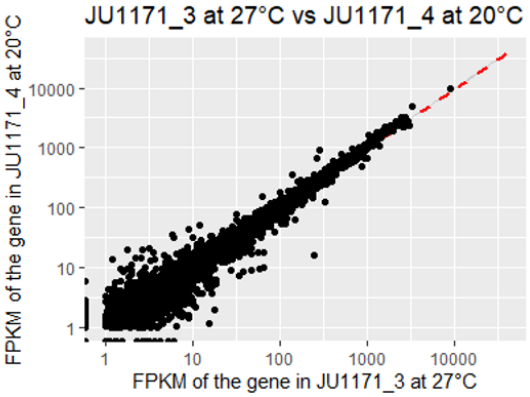

JU1171_2 at 27°C vs JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

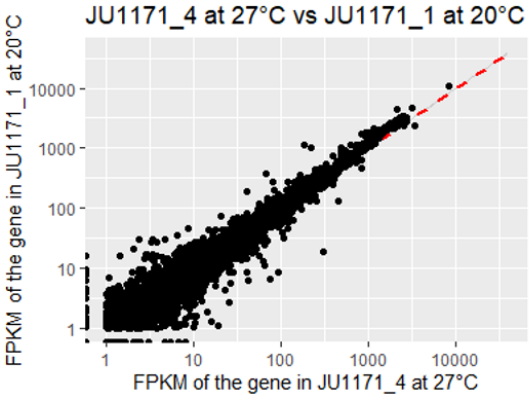

JU1171_2 at 27°C vs JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 20°C
FPKM of the gene in JU1171_2 at 27°C
1
10
100
1000
10000

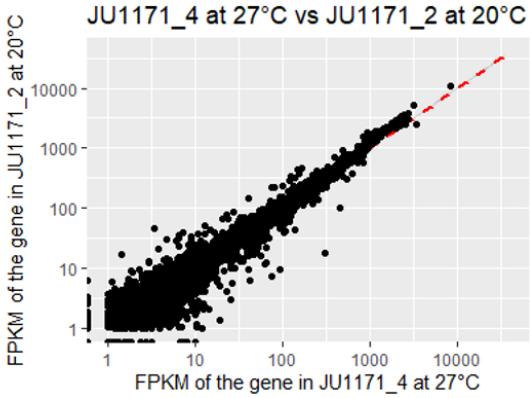

JU1171_3 at 27°C vs JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 20°C
FPKM of the gene in JU1171_3 at 27°C
1
10
100
1000
10000

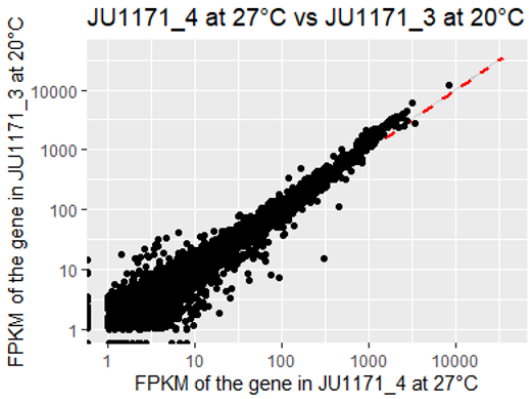

JU1171_3 at 27°C vs JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 20°C
FPKM of the gene in JU1171_3 at 27°C
1
10
100
1000
10000

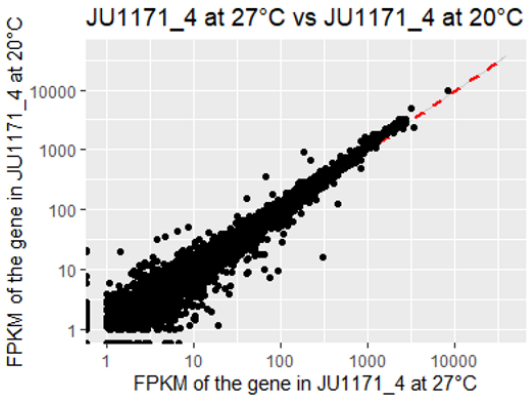

JU1171_3 at 27°C vs JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 27°C
1
10
100
1000
10000

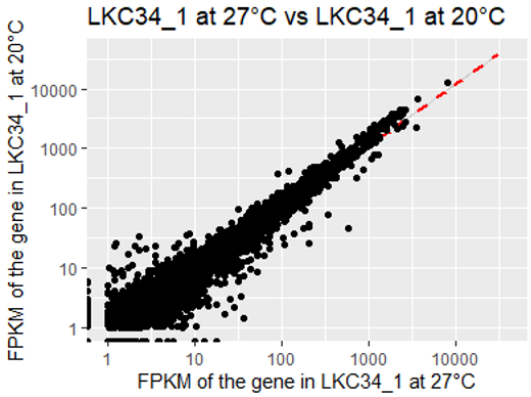

JU1171_3 at 27°C vs JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 20°C
FPKM of the gene in JU1171_3 at 27°C
1
10
100
1000
10000

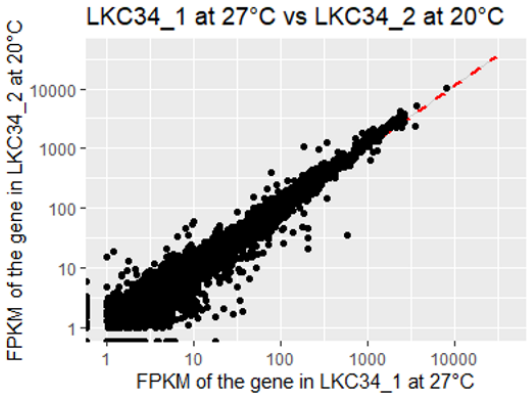

JU1171_4 at 27°C vs JU1171_1 at 20°C
FPKM of the gene in JU1171_1 at 20°C
FPKM of the gene in JU1171_4 at 27°C
1
10
100
1000
10000

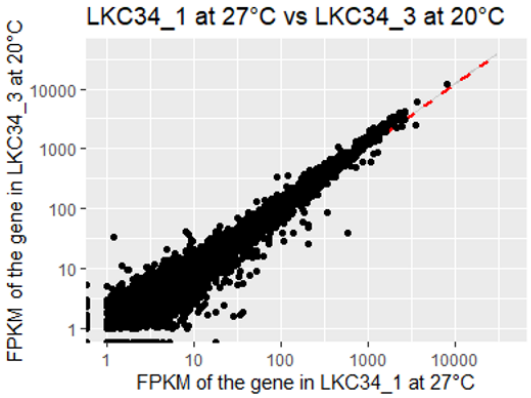

JU1171_4 at 27°C vs JU1171_2 at 20°C
FPKM of the gene in JU1171_2 at 20°C
FPKM of the gene in JU1171_4 at 27°C
1
10
100
1000
10000

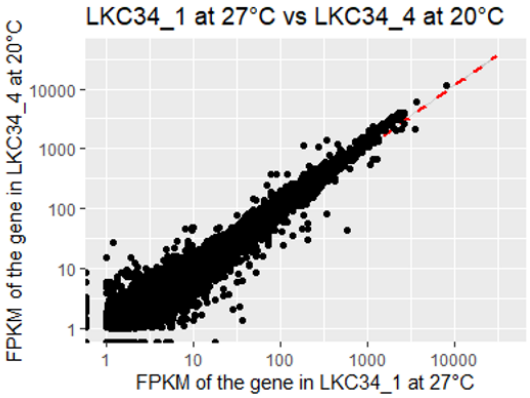

JU1171_4 at 27°C vs JU1171_3 at 20°C
FPKM of the gene in JU1171_3 at 20°C
FPKM of the gene in JU1171_4 at 27°C
1
10
100
1000
10000

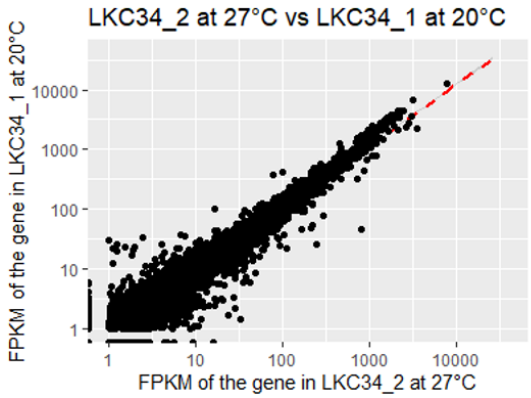

JU1171_4 at 27°C vs JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 20°C
FPKM of the gene in JU1171_4 at 27°C
1
10
100
1000
10000

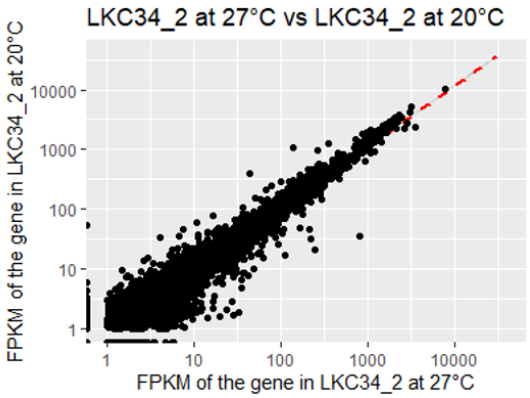

LKC34_1 at 27°C vs LKC34_1 at 20°C
FPKM of the gene in LKC34_1 at 20°C
FPKM of the gene in LKC34_1 at 27°C
1
10
100
1000
10000

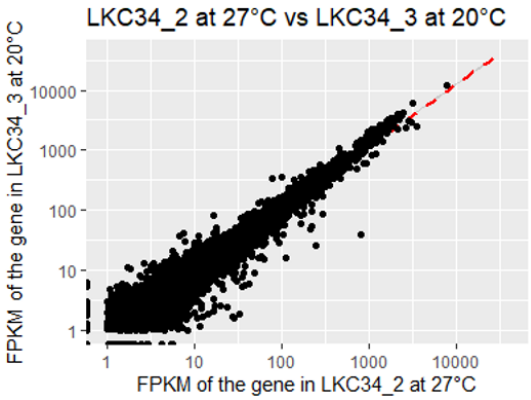

LKC34_1 at 27°C vs LKC34_2 at 20°C
FPKM of the gene in LKC34_2 at 20°C
FPKM of the gene in LKC34_1 at 27°C
1
10
100
1000
10000

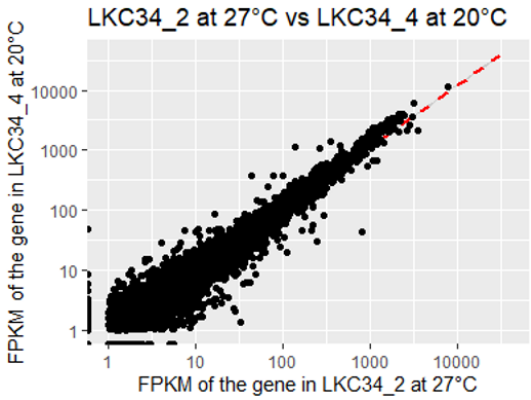

LKC34_1 at 27°C vs LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 20°C
FPKM of the gene in LKC34_1 at 27°C
1
10
100
1000
10000

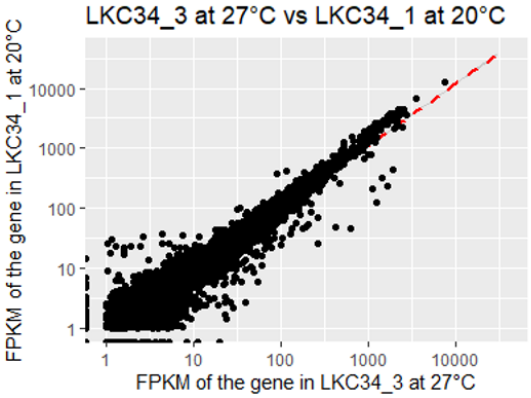

LKC34_1 at 27°C vs LKC34_4 at 20°C
FPKM of the gene in LKC34_4 at 20°C
FPKM of the gene in LKC34_1 at 27°C
1
10
100
1000
10000

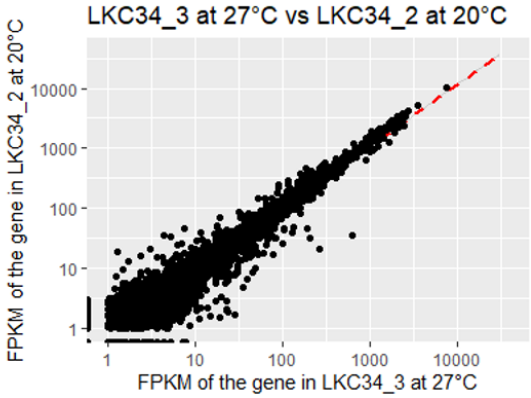

LKC34_2 at 27°C vs LKC34_1 at 20°C
FPKM of the gene in LKC34_1 at 20°C
FPKM of the gene in LKC34_2 at 27°C
1
10
100
1000
10000

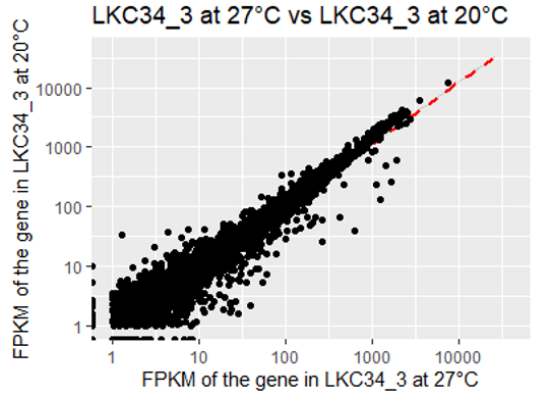

LKC34_2 at 27°C vs LKC34_2 at 20°C
FPKM of the gene in LKC34_2 at 20°C
FPKM of the gene in LKC34_2 at 27°C
1
10
100
1000
10000

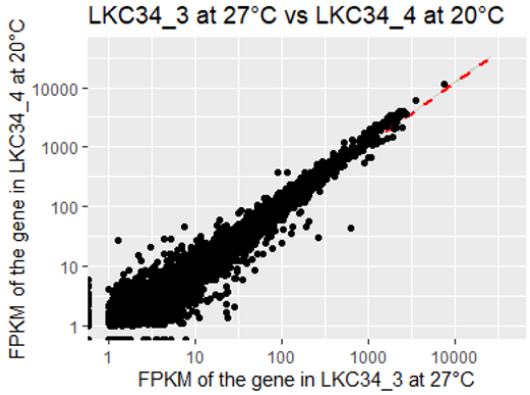

LKC34_2 at 27°C vs LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 20°C
FPKM of the gene in LKC34_2 at 27°C
1
10
100
1000
10000

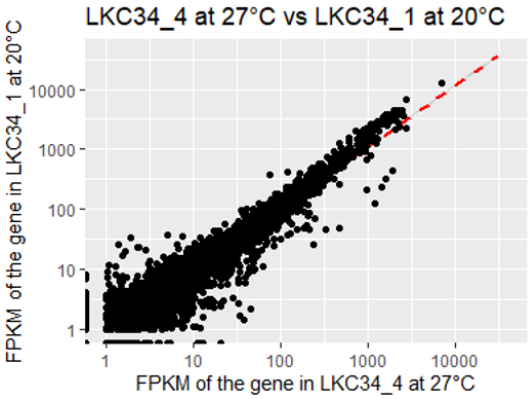

LKC34_2 at 27°C vs LKC34_4 at 20°C
FPKM of the gene in LKC34_4 at 20°C
FPKM of the gene in LKC34_2 at 27°C
1
10
100
1000
10000

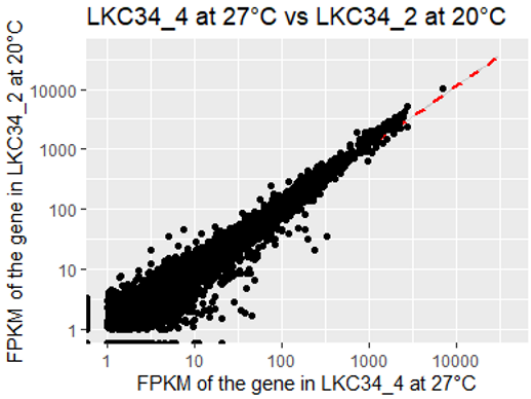

LKC34_3 at 27°C vs LKC34_1 at 20°C
FPKM of the gene in LKC34_1 at 20°C
FPKM of the gene in LKC34_3 at 27°C
1
10
100
1000
10000

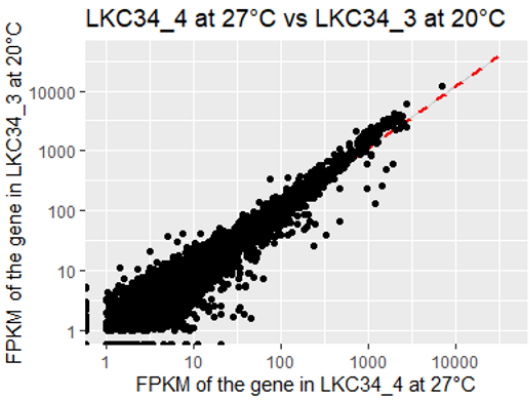

LKC34_3 at 27°C vs LKC34_2 at 20°C
FPKM of the gene in LKC34_2 at 20°C
FPKM of the gene in LKC34_3 at 27°C
1
10
100
1000
10000

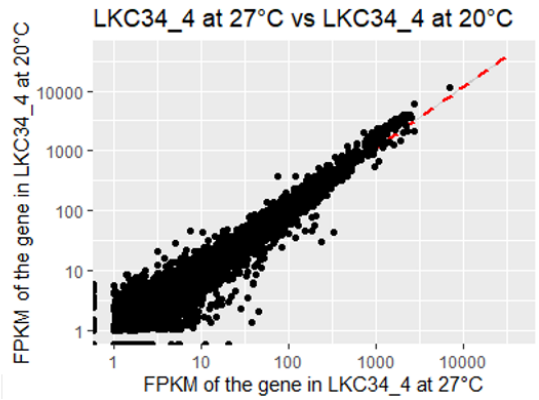

LKC34_3 at 27°C vs LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 20°C
FPKM of the gene in LKC34_3 at 27°C
1
10
100
1000
10000


LKC34_3 at 27°C vs LKC34_4 at 20°C
FPKM of the gene in LKC34_4 at 20°C
FPKM of the gene in LKC34_3 at 27°C
1
10
100
1000
10000


LKC34_4 at 27°C vs LKC34_1 at 20°C
FPKM of the gene in LKC34_1 at 20°C
FPKM of the gene in LKC34_4 at 27°C
1
10
100
1000
10000

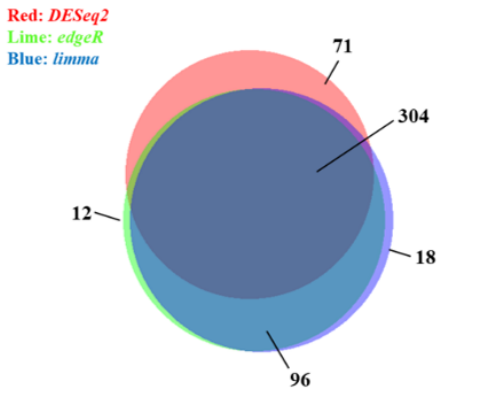

LKC34_4 at 27°C vs LKC34_2 at 20°C
FPKM of the gene in LKC34_2 at 20°C
FPKM of the gene in LKC34_4 at 27°C
1
10
100
1000
10000



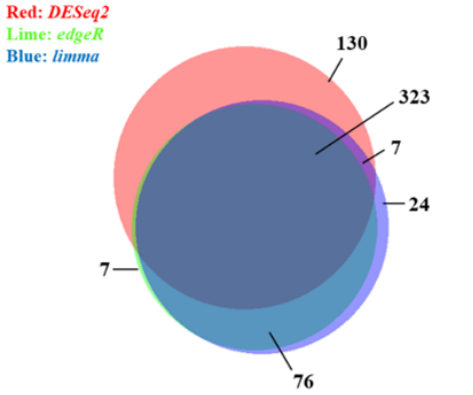


Scatter plots were created to compare the gene expression of each replicate among the JU1171 at 20°C versus 27°C, LKC34 at 20°C versus 27°C. The red dashed lines are the regression lines. X-axis and Y-axis are the FPKM values of the gene in the two replicates respectively. Axes were rendered on the log10 scale.

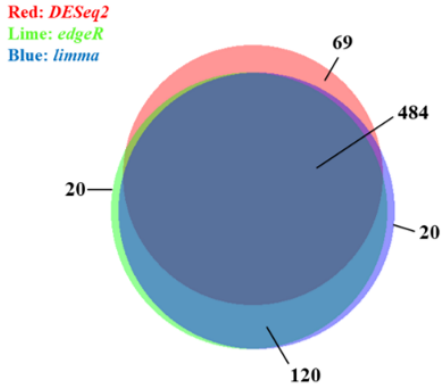
Appendix 3.



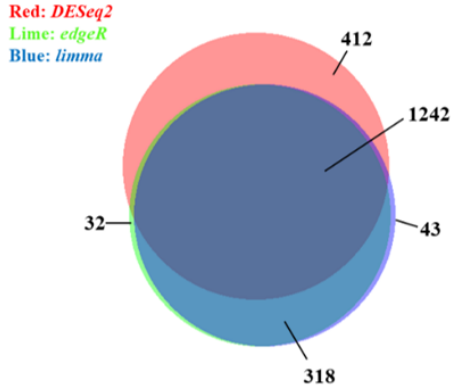
higher in JU1171 at 20°C than in LKC34 at 20°C



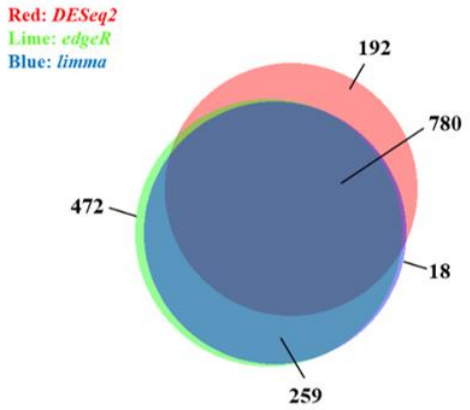
higher in LKC34 at 20°C than in JU1171 at 20°C



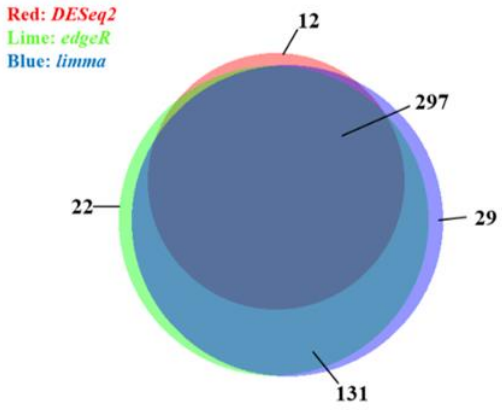
higher in JU1171 at 20°C than in LK34 at 27°C



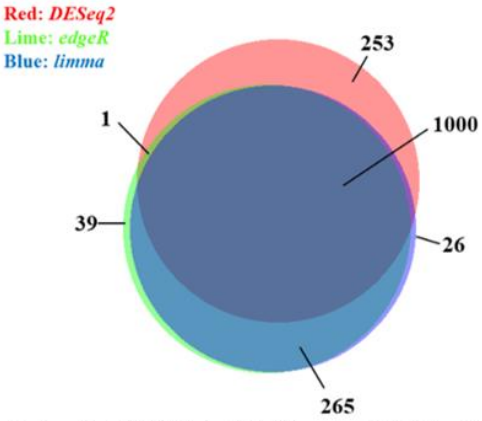
higher in LK34 at 27°C than in JU1171 at 20°C



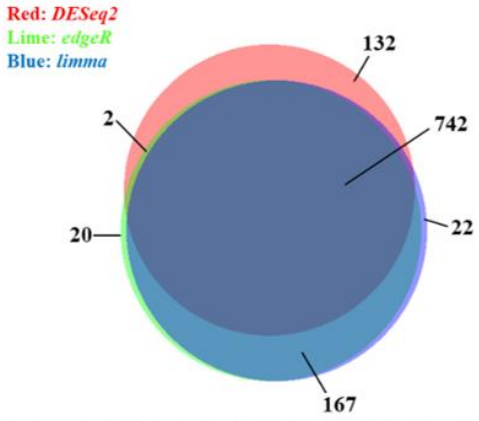
higher in JU1171 at 27°C than in JU1171 at 20°C



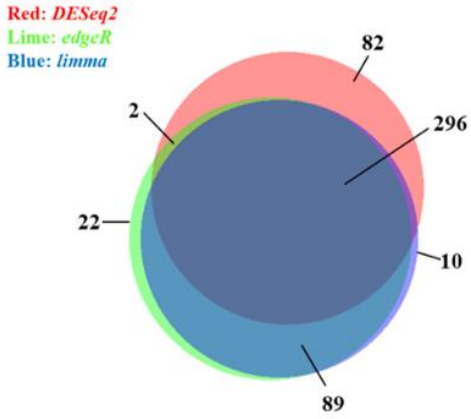
higher in JU1171 at 20°C than in JU1171 at 27°C



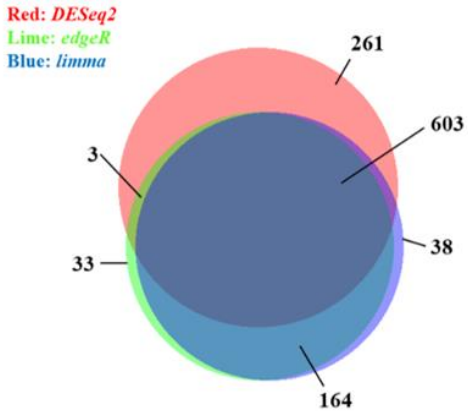
higher in JU1171 at 27°C than in LKC34 at 20°C



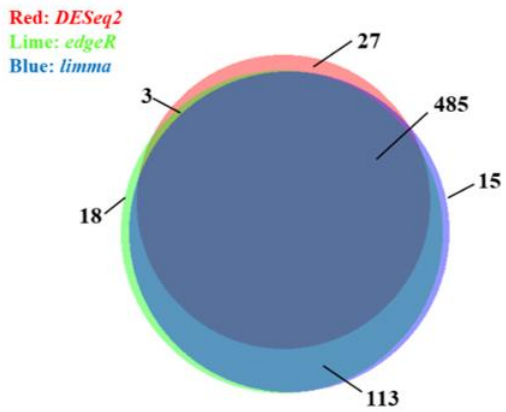
higher in LKC34 at 20°C than in JU1171 at 27°C



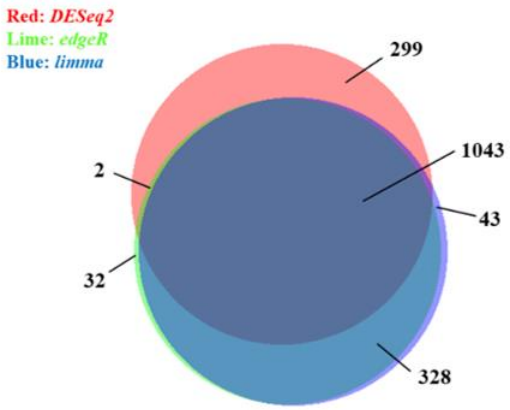
higher in JU1171 at 27°C than in LKC34 at 27°C



higher in LKC34 at 27°C than in JU1171 at 27°C



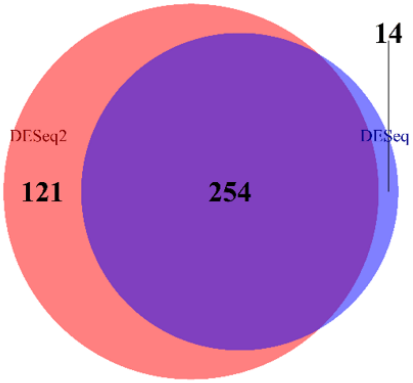
higher in LKC34 at 20°C than in LKC34 at 27°C



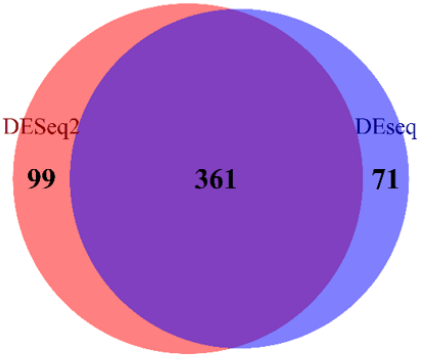
higher in LKC34 at 27°C than in LKC34 at 20°C

The Venn-diagrams were used to display the overlap among genes that were found to be significantly higher expressed under each condition by using the three differential expression testing tools: *DESeq2*, *edgeR* and *limma*.

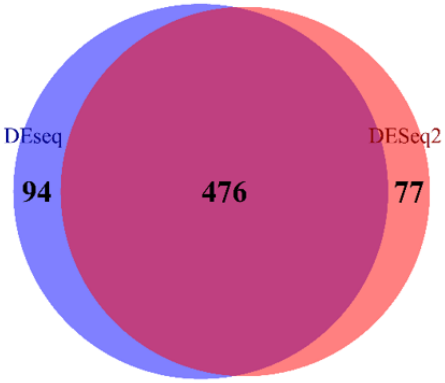
Appendix 4.



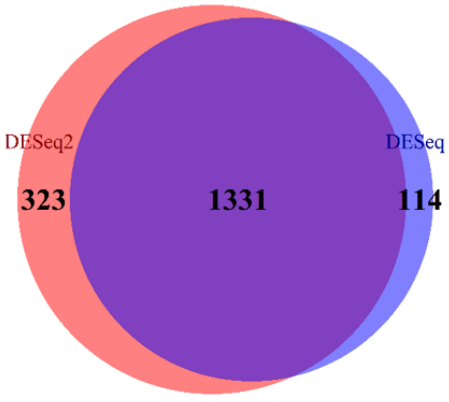
higher in JU1171 at 20°C than in LKC34 at 20°C



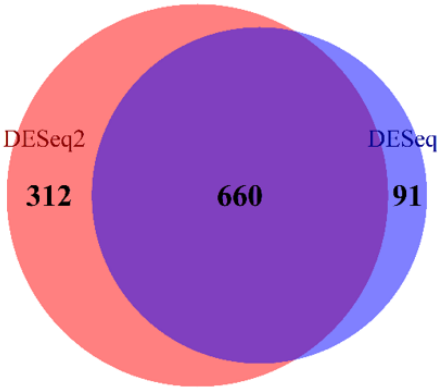
higher in LKC34 at 20°C than in JU1171 at 20°C



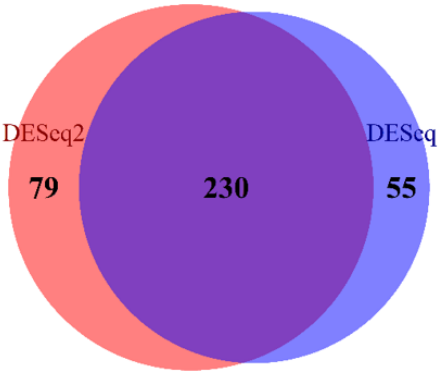
higher in JU1171 at 20°C than in LKC34 at 27°C



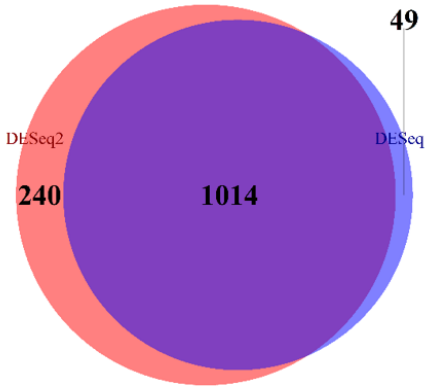
higher in LK34 at 27°C than in JU1171 at 20°C



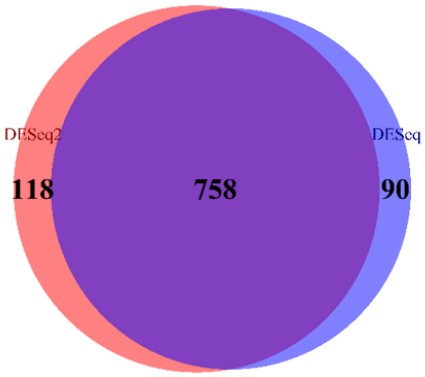
higher in JU1171 at 27°C than in JU1171 at 20°C



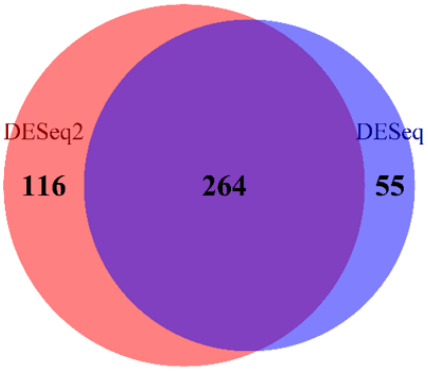
higher in JU1171 at 20°C than in JU1171 at 27°C



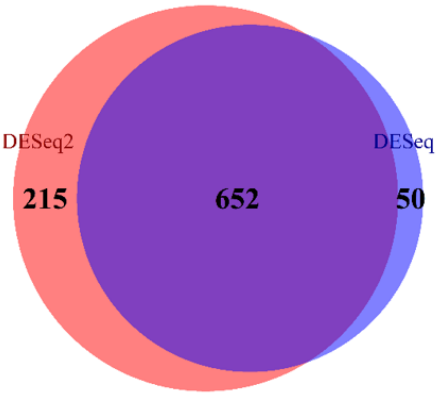
higher in JU1171 at 27°C than in LKC34 at 20°C



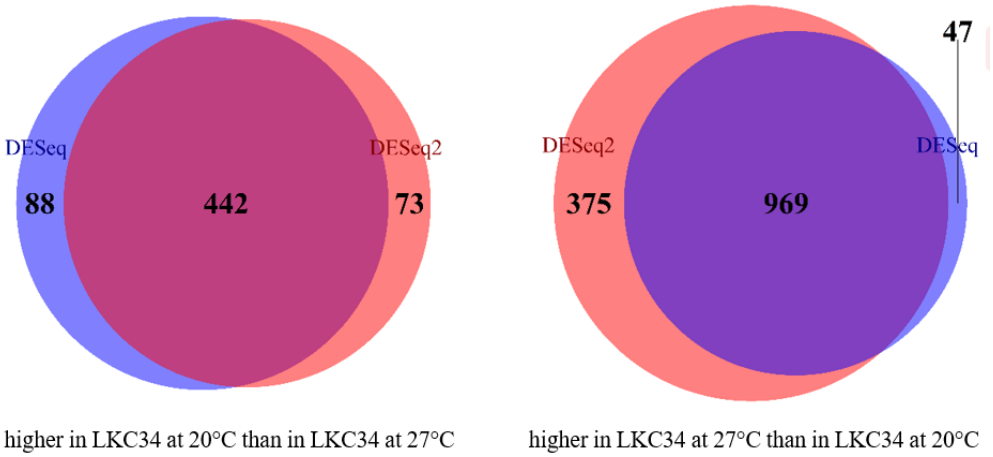
higher in LKC34 at 20°C than in JU1171 at 27°C



higher in JU1171 at 27°C than in LKC34 at 27°C



higher in LKC34 at 27°C than in JU1171 at 27°C



The Venn-diagrams were used to display the overlap among genes that were found to be significantly higher expressed under each condition by using the two differential expression testing tools: *DESeq* and *DESeq2*.