5-26-2016

Examining the evolutionary constraints of gene expression levels in S. Cerevisiae

Andrea Janelle Jackson
Rowan University

EXAMINING THE EVOLUTIONARY CONTSTRAINTS OF GENE EXRPESSION LEVELS IN *S. CEREVISIAE*

by

Andrea Jackson

A Thesis

Submitted to the
Department of Bioinformatics
College of Science and Mathematics
In partial fulfillment of the requirement
For the degree of
Master of Science in Bioinformatics
at
Rowan University
April 26, 2016

Thesis Chair: Dr. Mark Hickman

Dedications

I would like to dedicate this thesis to my husband, Dan, my mom, Gail, my dad, Joe, and my sister, Julie, who have always supported me in my academic ventures. Thank you for always encouraging me to do my best, but also reminding me to have fun and enjoy life too. And most of all, I dedicate this to God: Thank you for giving my life meaning and a purpose.

Acknowledgments

My sincerest thanks goes to Dr. Hickman who has helped me so much throughout this process. I am so thankful for his patience and continual encouragement. He taught me how to write code in R, showed me how to use the command line, proofread my papers, helped me with my research every week, and even provided delicious donuts at meetings! I couldn't have asked for a better research advisor!

I would also like to thank the members of my research team, Julianne Thorton, Abigail Smith, and Amanda Tursi, for all of their hard work in solving extensive computational problems, and Dr. Breitzman for teaching me how to use many awesome data mining techniques.

Abstract

Andrea Jackson
EXAMINING THE EVOLUTIONARY CONSTRAINTS OF GENE EXPRESSION LEVELS IN *S. CEREVISIAE*
2015- 2016
Mark Hickman, Ph.D.
Master of Science in Bioinformatics

It has been widely reported that some genes are expressed at a higher level than others. However, it has not been shown whether each gene is expressed consistently between studies and at the same level compared to all other genes. Here, we examined six RNA-seq datasets and found that the mRNA level of each gene is indeed consistent relative to all other genes. This result implies that there are evolutionary pressures that drive genes to maintain either low or high expression. In order to identify these pressures, we compared gene expression level to the features of each gene (or associated protein), such as biological function, molecular process, or localization. We found many possible pressures; for example, genes involved in translation and the ribosomal processes were expressed at high levels while genes involved in transcription and DNA-related processes were expressed at low levels. Furthermore, through the optimization of an artificial neural network, we were able to use several of these features to predict gene expression with 65-75% accuracy. In conclusion, these results show that gene expression level is controlled by several evolutionary constraints, including biological function, molecular process, and cellular localization.

Table of Contents

Abstract v

List of Figures viii

List of Tables x

Chapter 1: Introduction 1

The Evolution of Gene Expression 1

Expression in Eukaryotes 1

Gene Expression Levels 3

The Need/Cost Balance 5

Chapter 2: Methods 8

Data Collection 8

RNA and Protein Expression Datasets 8

Gene Ontology(GO) and Localization Datasets 9

Data Preprocessing and Normalization 10

RNA-seq Preprocessing 10

Protein Abundance Preprocessing 11

RPKM Normalization 11

Rank Normalization 12

Correlations 13

Data Merging 14

Bootstrapping 14

Normalized RNA-seq Datasets 14

Protein Dataset 15

Neural Network....16
Chapter 3: Results....20
Correlations and Plots of Normalized RNA-seq Datasets....20
Correlations and Plots of Protein Datasets....24
Correlations and Plots of Normalized RNA-seq Datasets with Other Data....25
Descriptive Categories Sorted by Mean....29
Bootstrapping....35
Localization Bootstrapping....35
GO Bootstrapping....38
Artificial Neural Network Results....44
Normalized RNA-seq Data....44
Protein Abundance Data....46
Chapter 4: Discussion....48
Correlation Analysis....48
Limitations in RNA-seq & Protein Abundance Measurements....49
Expression Trends....50
Conclusion....52
Future Research....53
References....54

List of Figures

Figure Page

Figure 1. Sample of artificial neural network structure with localization categories as the input and RNA level as the output 17

Figure 2. Expression values of genes in the upper (high) and lower (low) quartiles 18

Figure 3. Spearman correlation coefficients and plots for studies in the RPKM dataset 21

Figure 4. Spearman correlation coefficients and plots for studies in the low variability RPKM dataset 22

Figure 5. Spearman correlation coefficients and plots for studies in the Rank dataset 23

Figure 6. Spearman correlation coefficients and plots for studies in the low variability rank dataset 24

Figure 7. Spearman correlation coefficients and plots for protein abundance datasets 25

Figure 8. The natural log of average RPKM values (with low variability) was plotted against the natural log of the protein abundances values obtained by Chong et al. (2015), Newman et al. (2006), and Ghaemmaghami et al. (2003) 26

Figure 9. The ranks of RNA-seq values were plotted against ranks of protein abundances derived from the datasets by Chong et al. (2015), Newman et al. (2006), and Ghaemmaghami et al. (2003) 27

Figure 10. This figure shows the plots of normalized RNA-seq data against RNA abundance data normalized by rank *(left)* or transformed by natural log *(right)* obtained through affymetrix microarray by Legronne et al. (2004) 29

Figure 11. Statistically Significant Localization Categories in the RPKM Dataset 35

Figure 12. Statistically Significant Localization Categories in the Low Variability RPKM Dataset 36

Figure 13. Statistically Significant Localization Categories in the Rank Dataset 37

Figure 14. Statistically Significant Localization Categories in the Protein Dataset 38

Figure 15. Statistically Significant GO Categories in the RPKM Dataset 40

Figure 16. Statistically Significant GO Categories in the Low Variability RPKM Dataset 41

Figure 17. Statistically Significant GO Categories in the Rank Dataset 42

Figure 18. Statistically Significant GO Categories in the Protein Abundance Dataset 43

Figure 19. Prediction Accuracy of Artificial Neural Network for RNA-seq data 45

Figure 20. Prediction Accuracy of Artificial Neural Network for Newman et al. (2006) Protein Abundance data 47

List of Tables

Table Page

Table 1. Description of Expression Datasets 9

Table 2. RNA-seq Spearman's Rank Correlation Coefficients 28

Table 3. Comparison of RNA-seq and Protein Dataset means for Each Localization Category 30

Table 4. Comparison of RNA-seq and Protein Dataset Means for each GO Category ...31

Table 5. Average Prediction Accuracy of Artificial Neural Network for Normalized RNA-seq Datasets 45

Table 6. Average Prediction Accuracy of Artificial Neural Network for Neman et al. (2006) Protein Abundance Dataset 47

Chapter 1

Introduction

The Evolution of Gene Expression

It has been found that genes are expressed at varying levels (Rifkin, Kim, & White, 2003). Some are expressed at high levels, while others are expressed at low levels. For example, when measured at the same time and under the same conditions in yeast, the ribosomal protein encoding gene, YDL133C-A, has been shown to be expressed at high levels, while the membrane protein encoding gene, YBR04W, has been shown to be expressed at low levels (Fox et al., 2015; Jenner et al., 2012; Erdman et al., 1998). In the yeast genome, approximately 20 percent of protein coding genes are expressed at high levels, while 33 percent are expressed at low levels (Nagalakshmi et al., 2008).

Researchers who have studied the evolution of expression levels have shown that the expression levels of different types of organisms cluster based on tissue type and are somewhat conserved across related species (Brawand, et al., 2011; Nuzhdin et al., 2004). However, research is lacking on the topic of the evolution of steady-expression levels in cells. In other words, what evolutionary pressures would cause some genes, in consistent conditions, to be always expressed at high levels, while others are always expressed at low levels? Here, we attempt to understand the evolutionary pressures that influence the level of gene expression.

Expression in Eukaryotes

"Gene expression" is the process in which the information in a gene, a segment of deoxyribonucleic acid (DNA), is used to create a protein, the molecular machinery of the cell (Booth & Lees, 2007). Gene expression consists of two sub-processes, transcription and translation. During transcription, gene information encoded in DNA, a double-

stranded macromolecule, is essentially copied into a single-stranded messenger ribonucleic acid (mRNA) (Krishnamurthy, S., Hampsey, M., 2009). The information in an mRNA molecule is used to generate a unique protein in translation, the second part of gene expression (Cooper, G. 2000; Hannan et al., 2003; Pestova et al., 2001). In *Saccharomyces cerevisiae,* the organism chosen for this study, approximately 6000 genes are used to create proteins (SGD Project, 2016b).

It is important to recognize that, while transcription and translation promote protein production and increase protein levels, mRNA degradation and protein degradation are two cellular processes that lead to lower protein levels (Schwanhäusser et al., 2011). mRNA degradation is a process that results in the decay of an mRNA molecule. While all mRNA molecules will eventually decay, the half-lives of mRNAs are variable. Generally, the mRNAs with the shortest half-lives are those that code for regulatory proteins. In *S. cerevisiae,* mRNA degradation assists the cell by maintaining the steady-state levels of mRNA expression, removing abnormal mRNAs, and removing unneeded RNA fragments (Cooper, 2000; Jackowiak, et al., 2011; Parker, 2012).

Similarly, protein degradation is a process that results in the decay of a protein macromolecule. Much like the rates of decay of mRNAs, the rates of decay of each type of protein is variable. Generally, regulatory proteins have the fastest decay rates. In cells, protein degradation provides a way for the cell to recycle proteins involved in cell regulation and to destroy abnormally folded or damaged proteins by breaking them down to amino acids. These amino acids can be used to form new proteins in the translation process (Cooper, 2000; Goldberg, 2003).

The processes of transcription, translation, mRNA degradation, and protein degradation all contribute to the levels of gene expression in a cell. However, all contributions are not equal. In their study on mammalian cells, Li, Bickel, and Biggin (2014) estimated that the greatest contribution to protein levels in a cell is determined by transcription (~38 - 73%) and followed by translation (~8 - 30%), RNA degradation (~11 - 18%), and protein degradation (~8 - 14%). While protein abundance levels are ideal measures of gene expression in a cell, obtaining accurate quantifications of these levels presents a challenge for researchers.

Gene Expression Levels

Some studies have used protein abundane4ce measurements to determine the levels of gene expression in cells (Albert et al., 2014; Yu et al., 2006). However, various limitations to the quantification of protein levels have been reported. When quantifying protein levels with mass spectrometry (MS) and two-dimensional gel electrophoresis, proteins that are more abundant in the cell tend to be more easily detected, resulting in inaccurate quantifications of proteins that are expressed at low levels. The quantification of protein levels though western blot analysis of epitope-tagged proteins has been shown to be more accurate for the detection of proteins expressed at low levels. However, the tag may interfere with proper regulation of the protein and thus may result in inaccurate abundance measurements (Ghaemmaghami et al., 2003). Hence, protein abundances cannot be accurately and consistently measured at the current time (Guimaraes, Rocha & Arkin, 2014).

As a result, many studies have used mRNA abundance as a proxy for protein abundance. While microarrays have been primarily used to study the differential

expression of mRNAs, RNA sequencing (RNA-seq), a more recently developed sequencing method has been often used for the analysis of genome-wide expression (Nookaew, et al., 2012). In this method, RNA is converted to cDNA, which is then fragmented, sequenced, and mapped to a genome. Due to its high sensitivity, RNA-seq is thought to be more effective for quantifying the expression of genes expressed at low levels, genes expressed at high levels, and closely related genes than the previously-used DNA microarrays (Mortazavi, et al., 2008a, Nagalakshmi et al., 2008, Wang et al., 2009). Additionally, because the RNA-seq method does not require hybridization to a probe, as required by the DNA microarray method, absolute gene expression levels can potentially be compared within a sample (Fu, et al., 2009; Hackett, et al., 2012; Marioni et al., 2008; Fu et al., 2009). However, measuring mRNA levels as a proxy for protein levels has limitations as well. Posttranscriptional events, like translation rate and protein degradation, are known to influence protein levels (Guimaraes, Rocha & Arkin, 2014).

While studying expression, many researchers have correlated mRNA levels with protein levels. Some studies have shown poor correlations between mRNA abundance and protein abundance, possibly indicating the importance of posttranscriptional events in controlling protein levels. However, these studies generally use a small number of representative genes, correlate mRNA and protein data obtained from different laboratories or studies, or leave out important factors such as protein or RNA decay (Schwanhausser et al., 2011). Recent studies have shown higher correlations between mRNA and protein than what had been shown previously. In a parallel study of the developing maize leaf, mRNA counts, obtained through RNA-seq, were shown to explain 40-70% of protein abundance measurements, obtained through label-free high resolution

mass spectrometry. Abundances corrected by length using the normalized spectral abundance factor (NSAF) for protein and reads per kilobase of exon model per million mapped reads (RPKM) for mRNA yielded higher correlations than abundances that were not length-corrected (Ponnala, et al., 2014). Similarly, in a parallel study of mRNA and protein abundance in mammalian cells (HeLa & NIH3T3), mRNA abundance was shown to explain 56-84% of protein abundance. This high percentage was found after all protein abundances were corrected with a regression model formed using a sample of protein abundance measures from reliable housekeeping genes (Li, et al., 2014). These studies show that relative mRNA levels can be used as an approximate representation of relative protein levels in the cell.

In recent years, the prevalence of genome-wide expression data has increased dramatically. A variety of public repositories have been created to store the ever growing collection of expression data. Because many journals now encourage researchers to submit their expression data to public repositories before their research articles are published, a large amount of high quality data is available in these data warehouses (Rung & Brazma, 2012). In this study, we took advantage of the high quality RNA-seq and protein expression data available in data repositories to study expression levels in *S. cerevisiae.*

The Need/Cost Balance

Trade-offs, a fundamental concept in evolutionary biology, occur when organisms are subject to limited resources (Stoebel et al., 2008). These resources vary widely from external sources such as food, water, or space to internal sources such as energy. In an organism, a limited store of energy must be distributed among all processes, which

results in a continual trade-off between energy costs and processes required by the organism (Niven & Laughlin, 2008). Transcription and translation are known to be energetically costly processes. In *S. cerevisiae*, increases in gene expression levels have been shown to be accompanied by increases in energy cost (Wagner, 2005). Therefore, an increase in the expression level of one gene affects the store of energy budgeted for the expression of other genes (Wagner, 2007).

It is likely that the wide variation in expression levels within a cell results from the balance between protein production cost and cellular need for a specific protein (Guimaraes, Rocha, & Arkin, 2014). In standard growth conditions (rich medium, 30°C) some *S. cerevisiae* genes have been shown to be expressed at high levels, while others have been shown to be expressed at mid or low levels (Ghaemmaghami et al., 2003; Newman et al., 2006). In the current study, we did not seek to identify *how* genes are expressed at high or low levels (e.g., regulation of transcription or translation). Rather, we sought to identify *why* genes are expressed at varied levels (e.g., proteins of a certain function are required at higher levels). To explore this phenomenon, we explored whether gene features may act as evolutionary pressures to influence the expression level of genes. Gene features were gathered from the GO (gene ontology) Slim Mapping dataset (SGD Project, 2016a), which attempts to characterize each gene product in terms of its biological processes, molecular functions, and cellular components. Additionally, we used the Yeast GFP Fusion Localization database (University of California Regents, 2006), which contains the names of 22 possible subcellular localization compartments and identifies the yeast proteins found in each of those compartments. Using a combination of mRNA and protein expression data, we investigated the processes,

components, functions, and localizations of proteins in the cell that could act as evolutionary pressures for the high or low expression of genes in *S. cerevisiae.*

Chapter 2

Methods

Data Collection

RNA and protein expression datasets. In order to study the variation in gene expression, RNA-seq datasets were collected from six studies, using the similar yeast strains and growth conditions (Adhakari & Cullen, 2014; Baker et al., 2013; Fox et al., 2015; Hickman, 2016; Martin et al., 2014; Risso et al., 2011). All yeast cells were grown in yeast extract peptone dextrose (YPD), either pure or with the addition of dextrose, at temperatures ranging from 25 to 30 degrees Celsius, when specified. The wild type *S. cerevisiae* strains used in these studies were S288C, BY4741, a closely-related derivative of S288C (Brachmann et al., 1998), or Sigma 1278B. Additionally, the data from a study that quantified RNA levels through Affymetrix microarray methods was gathered for comparison with RNA-seq values (Lengronne et al., 2004). Likewise, studies that quantified protein abundance data through various methods were also gathered for comparison. A summary of the conditions and methods used in each individual study is in Table 1 (Chong et al., 2015; Ghaemmaghami et al., 2003; Newman et al., 2006).

Table 1

Description of Expression Datasets

	Strain	Conditions	Number of samples obtained from study	Number of genes obtained from study	Expression method	Platform	GEO# or SRA#	Reference
RNA	Sigma 1278B	YEPD, 30°C	3	5552	RNA-seq	Illumina HiSeq 2500	GSE61783	Adhikari & Cullen, 2014
	BY4741	YPD	3	6691	RNA-seq	Illumina HiSeq 2000	GSE43002	Baker, et al., 2013
	BY4741	YPD	4	5750	RNA-seq	AB 5500xl Genetic Analyzer	GSE57155	Fox et al, 2015
	BY4741	YPD, 30°C	1	6691	RNA-seq	Illumina HiSeq 2000	GSE52086	Martin et al., 2014
	S288C	YPD, 30°C	1	7130	RNA-seq	Illumina HiSeq		Hickman, unpublished
	S288C	YPD, 25°C	6	6691	RNA-seq	Imina Genome Analyzer II	SRA048710	Risso et al., 2011
	BY4741	YP with 2% glucose	1	6457	Affymetrix Microarray	[YG_S98] Affymetrix Yeast Genome S98 Array	GSM24746	Lengronne, et al., 2004
Protein	BY4872	low fluorescence synthetic medium with methionine, NAT, and 2% glucose	3	3540	Synthetic Genetic Array			Chong et al., 2015
	BY4741	SD	1	3868	Microscopy	Nikon TE200/300		Ghaemmaghami et al., 2003
	BY4741, BY4742	YEPD, 30°C	1	2385	High Throughput Flow Cytometry			Newman et al., 2006

Gene ontology (GO) and localization datasets. In order to explore potential evolutionary pressures, datasets containing information about yeast gene ontology (GO) and localization were collected. The gene ontology data was collected from the GO Slim

Mapping dataset (SGD Project, 2016). The GO dataset identifies whether gene products are aspects of the yeast cell's biological processes, molecular functions, or cellular components (The Gene Ontology Consortium, 2005). Non-coding genes, including those identified as tRNA, rRNA, ncRNA, snRNA, or snoRNA, were removed from the GO dataset. Additionally, GO categories that could bias the results in the further analysis were removed from the GO dataset. These categories were "cellular component", "molecular function", "biological process", "other", and "not yet annotated". The localization data was collected from the Yeast GFP Fusion Localization database (University of California Regents, 2006). This set contains the names of 22 possible subcellular localization compartments and identifies the yeast proteins found in each of those compartments (Huh et al., 2003). One localization category entitled "ambiguous" was removed from the dataset.

Data Preprocessing and Normalization

RNA-seq preprocessing. RNA-seq datasets were preprocessed to prepare for further analyses. Sets that were published in SRA format (Baker et al., 2013; Risso et al., 2011; Martin et al., 2014) were converted to FASTQ with the SRA toolkit (Leinonen, Sugawara, and Shumway, 2011), trimmed with the FASTQ quality trimmer using a quality score of ten (Blankenberg et al., 2010), and mapped to the most complete and recent *S. cerevisiae* genome build, R64.1.1 2011-02-03 (Engel et al., 2014), with TopHat (Kim et al., 2013). The number of sequencing reads per gene was calculated using the HTSeq program (Anders, Pyl, and Huber, 2015), generating a raw count file. Datasets that were published in raw count format were used in their published format (Adhikari, et al., 2014; Hu et al., 2014; Li et al., 2014; Fox et al., 2015). Some datasets used a

combination of common and systematic gene names and were converted to contain only systematic gene names through the use of the Saccharomyces Genome Database (SGD) (Cherry et al., 2012). Similarly, the reference IDs in the Affymetrix microarray RNA dataset by Legronne et al. (2004) were converted to systematic gene names.

Protein abundance preprocessing. The protein abundance data for each gene, collected from Chong et al. (2015), Ghaemmaghami et al. (2003), and Newman et al. (2006), was averaged to form the protein abundance dataset. This dataset was comprised of the systematic gene identification name for each protein associated gene, along with its corresponding average protein abundance.

Later, in the bootstrapping and artificial neural network analyses, the protein abundance dataset was simplified to include only data from the Newman et al. (2006) dataset.

RPKM normalization. After pre-processing was complete, two datasets were created by different normalization methods. The first normalization method used was a modified reads per kilobase of transcript per million mapped reads (RPKM) normalization. With this method, the RNA-seq counts for each gene were normalized with a variation of the RPKM formula the paper by Mortazavi et al. (2008b). The formula used by Mortazavi et al. (2008b) is as follows:

$$R = \frac{10^9 C}{NL}$$

$C =$ the number of reads that were mapped to a gene
$N =$ total number of reads that were mapped in the experiment
$L =$ total number of base pairs in the gene's exon

Gene exon length (L) was determined through the use of the SGD features file (SGD Project, 2015). The variation in our normalization methods occurred in the variable

N. Because our data included the raw counts from multiple experiments, we took the total number of mappable reads in genes which were common to all experiments and used that number as the representation of total mappable reads. The resulting RPKM values for each gene in all samples were averaged. The final RPKM dataset was comprised of the systematic name of each gene, along with the corresponding average RPKM value.

It is possible that some genes showed higher variability in counts due to differences in the strains used, slight differences in conditions, or errors in sequencing. The genes which showed the most consistency across experiments were kept in the Low Variability RPKM dataset for further analysis. These genes were selected by removing the upper quartile of genes which contained the highest standard deviation and keeping 75% of the genes with the lowest standard deviation. The final Low Variability RPKM dataset was comprised of the systematic gene identification name for each gene, along with its corresponding average RPKM value.

Rank normalization. The second normalization method used was a rank normalization approach. In this method, the genes of each study were ranked by expression level. If the RPKM or fragments per kilobase of transcript per million mapped reads (FPKM) were calculated by the researchers of the study, these values were used. If only the raw counts were available, we calculated RPKM for each sample within each study using the RPKM formula above without the "N" modification. In each study, the genes in each sample were assigned a number from 1 to n (where n represents the number of genes in a sample). Genes with lower RPKM/FPKM values were given lower ranks than those with high values. If two genes in a sample had the same value, they were assigned the same rank. Genes with an RNA-seq count of "0" were excluded from the

rankings. The ranks from each sample and corresponding experiment were merged by gene name to create a dataset with all of the samples. Genes with less than three observations throughout all samples were excluded from further analysis. Because the value of n varied from sample to sample, the ranks in each sample were normalized with the formula:

$$r - (1 - \frac{n_S}{n_L})$$

$r =$ rank of the gene

$n_S =$ number of genes in the sample

$n_L =$ number of genes in the largest sample

After normalization, the average rank was calculated for each gene, by averaging the ranks from all samples in all studies. The final rank dataset was comprised of the systematic gene identification name for each gene, along with its corresponding average rank value.

A set with low variability was comprised with the same methods that were used to comprise the low variability RPKM dataset. This dataset was named the low variability rank dataset.

Correlations

The correlations of samples within each RNA-seq dataset were calculated using the Spearman's rank correlation (r_s) method. This method was chosen because it can be used to correlate two sets of data that are not normally distributed (Ruscio, 2008). The Spearman's rank correlation method was also used to correlate affymetrix microarray values with average RPKM and average rank values, to correlate protein samples against other protein samples, to correlate protein samples with average RPKM and average rank values, and to correlate degradation datasets with protein abundance and normalized

RNA-seq datasets. The Spearman's rank correlation coefficients and their corresponding plots are shown in the results section.

Data Merging

Average RPKM and average rank datasets were merged separately with GO and localization datasets for further analysis. These sets were merged by systematic gene name. Each merged set contained a column of systematic gene names, a corresponding column of normalized RPKM or rank values, and a corresponding column of GO or localization categorizations.

The protein abundance set from Newman et al. (2006) were also merged with GO and localization sets using the same methods.

Bootstrapping

Normalized RNA-seq datasets. To determine which categories in the GO and localization datasets had statistically significant distributions, a bootstrapping program was written. The input for this program consisted of the RNA-seq averages from the normalized RPKM or rank datasets which were merged with the corresponding genes in the GO or localization categories. The mean of normalized RNA-seq averages in each individual GO or localization category was compared to the mean of randomly selected RNA-seq values from other categories ten thousand times. The number of random RNA-seq values selected for each category comparison equaled the number of genes in the category. If the mean of the actual distribution was higher than the mean of the distribution of randomly selected genes, the comparison was assigned the number one. If the mean of the actual distribution was lower than the mean of the distribution of randomly selected genes, the comparison was assigned the number zero. The ones and

zeros resulting from the ten thousand comparisons were added to a vector containing a number one (the number 1 prevents a p value of 0). This vector sum was used to determine the p value of the corresponding category. The p value was calculated through a two sided test, using the following formula, where n represents the number of comparisons:

$$p = \frac{\left|\left|\left(\frac{n}{2}+1\right) - vector\ sum\right| - \left(\frac{n}{2}+1\right)\right|}{n}$$

For this study, n was equal to ten thousand comparisons. To address multiple hypothesis testing, the calculated p values were adjusted with the p.adjust function in R (R core team, 2015) using the Benjamini & Hochberg (also called "false discovery rate") method. To validate the method, bootstrapping was also conducted by collecting the sums of the distributions, rather than the means.

Protein dataset. In order to further validate the use of RNA-seq bootstrapping results, bootstrapping was conducted on the Newman et al. (2006) protein dataset. Out of the three protein datasets discussed in this report, this protein dataset was chosen for the bootstrapping analysis because it contained protein abundance data obtained from strains and conditions most similar to the RNA-seq datasets. The bootstrapping analysis for this protein dataset was conducted in the same manner as the RNA-seq datasets, with one exception. If the actual mean was equal to the random mean in the bootstrapping statistical program, 0.5 was added to the sum vector. This step was not necessary for the bootstrapping analysis done on the RNA-seq datasets because equivalent values would be extremely rare. However, replicate abundance values in the protein dataset along with the smaller size of the protein dataset increase the appearance of equivalent values in the bootstrapping analysis.

Neural Network

A neural network was used to detect the presence of hidden patterns in the GO and localizations datasets. If distinct gene expression patterns exist, the GO and localization categories should be able to predict genes expressed at high or low levels in the RPKM, rank, and protein abundance (Newman et al., 2006) datasets. The neural network was constructed using neuralnet, an R package developed by Günther and Fritsch (2010). This neural network was initially composed of one hidden layer, three nodes, and a threshold of 0.005 and used resilient backpropagation with weight backtracking. It was later optimized to contain one hidden layer and two nodes. A sample of the neural network structure is shown in figure 1. Because new weights were generated each time a neural network repetition was run, the weights shown in the figure are not representative of the weights in all repetitions.

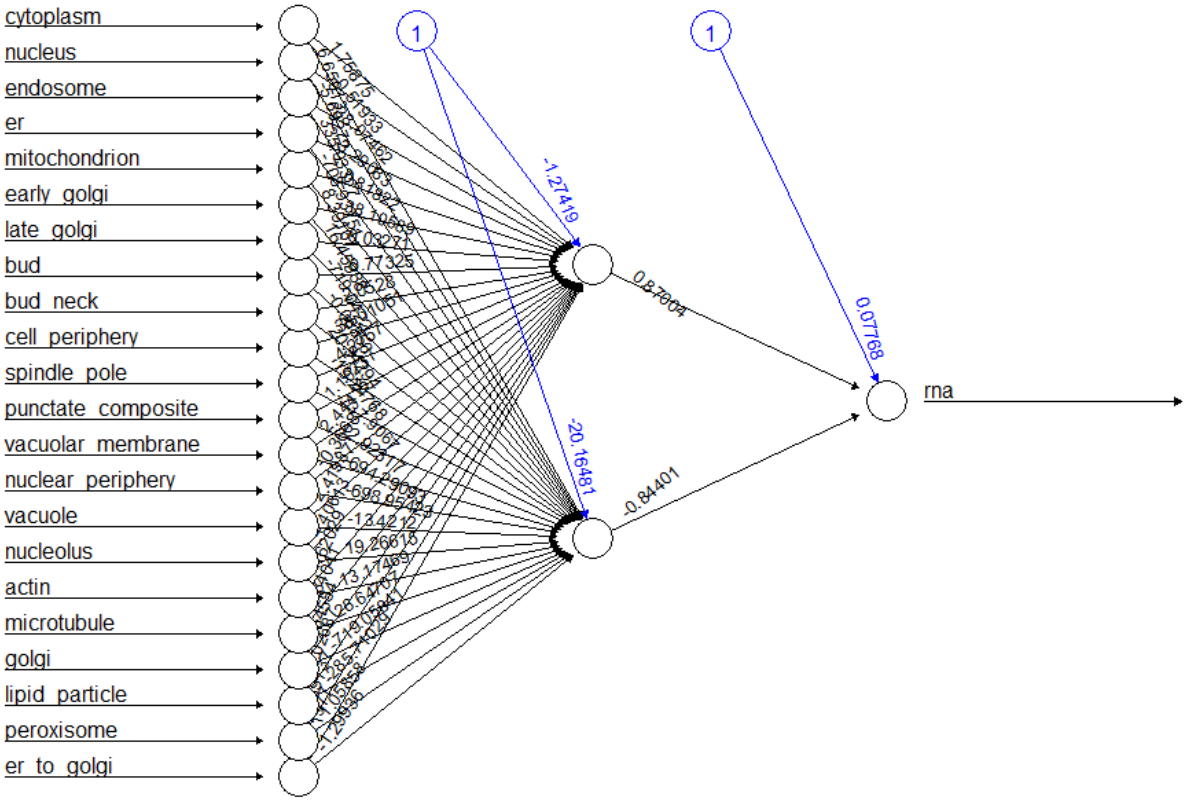

Figure 1. Sample of artificial neural network structure with localization categories as the input and RNA level as the output

GO and localization categories were converted into a binary matrix for use in the neural network. A "1" in a GO or localization category indicated that a gene belonged to that category; a "0" indicated that it did not.

RPKM, rank, and protein abundance values were normalized for the neural network using min/max normalization, where the minimum was 0 and the maximum was 1 (Patro & Kumar sahu, n.d.). In order to reduce noise, the neural network was only used for the quartiles of genes with the highest and lowest expression (see example in figure 2). Genes in the upper quartile were assigned a value of "1" and genes in the lower quartile were assigned a value of "0".

For each dataset combination, the neural network was trained and tested on data through 50 repetitions. Each repetition analyzed a random split of data, with eighty percent of the data used as training data and twenty percent used as testing data.

Random binary vectors were used to test the significance of the neural network results. These vectors were created to have the same length of the RPKM, rank, and protein abundance binary vectors, but were composed of randomly generated binary values, rather than upper/lower quartile binary generated values. The neural network tested the ability of the GO and localization categories to predict the random binary vectors with the same methods that were used to predict the RPKM, rank, and protein abundance binary vectors.

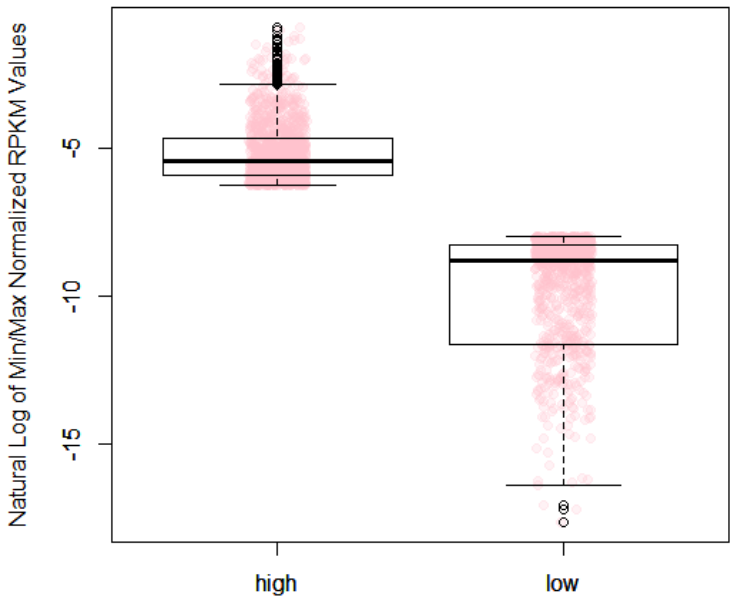

Figure 2. Expression values of genes in the upper (high) and lower (low) quartiles.

The prediction accuracy of each repetition of the neural network was determined by the following formula:

$$\frac{\sum_{i=1}^{n} TP + \sum_{i=1}^{n} TN}{\sum_{i=1}^{n} TP + \sum_{i=1}^{n} TN + \sum_{i=1}^{n} FP + \sum_{i=1}^{n} FN}$$

$TP =$ True Positive: Gene was correctly printed as high (1)
$TN =$ True Negative: Gene was correctly predicted as low (0)
$FP =$ False Positive: Gene was incorrectly predicted as high (1)
$FN =$ False Negative: Gene was incorrectly predicted as low (0)

The prediction accuracy for each repetition of the neural network was recorded and plotted.

Chapter 3

Results

Correlations and Plots of Normalized RNA-seq Datasets

In order to determine whether the expression level of each *S. cerevisiae* gene was consistent across several experiments, in which cells were grown under standard rich-media conditions, normalized RNA-counts from each study were correlated. The results of the correlations and plots for each normalized RNA-seq dataset show not only that the RNA-seq method is highly reproducible but that the expression level of each gene does not vary across experiments. In the average RPKM dataset, the averages of the samples from each study were compared to the averages of the samples from every other study. The Spearman correlation coefficients ranged from 0.684 to 0.964 (see Figure 3). The Spearman correlation coefficients resulting from the pairwise correlations of average RPKM values for studies in the low variability set ranged from 0.766 to 0.966 (see Figure 4). The p values of these all of these Spearman's rank correlations were $< 2.2\ e^{-16}$.

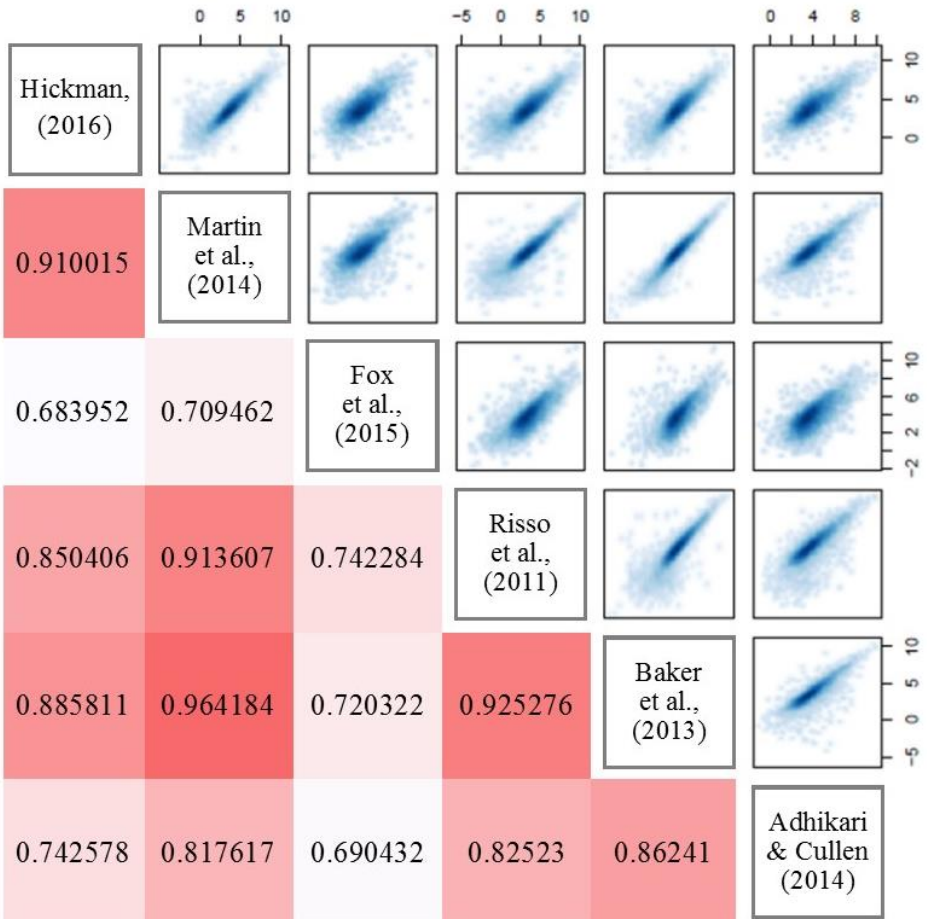

Figure 3. Spearman correlation coefficients and plots for studies in the RPKM dataset

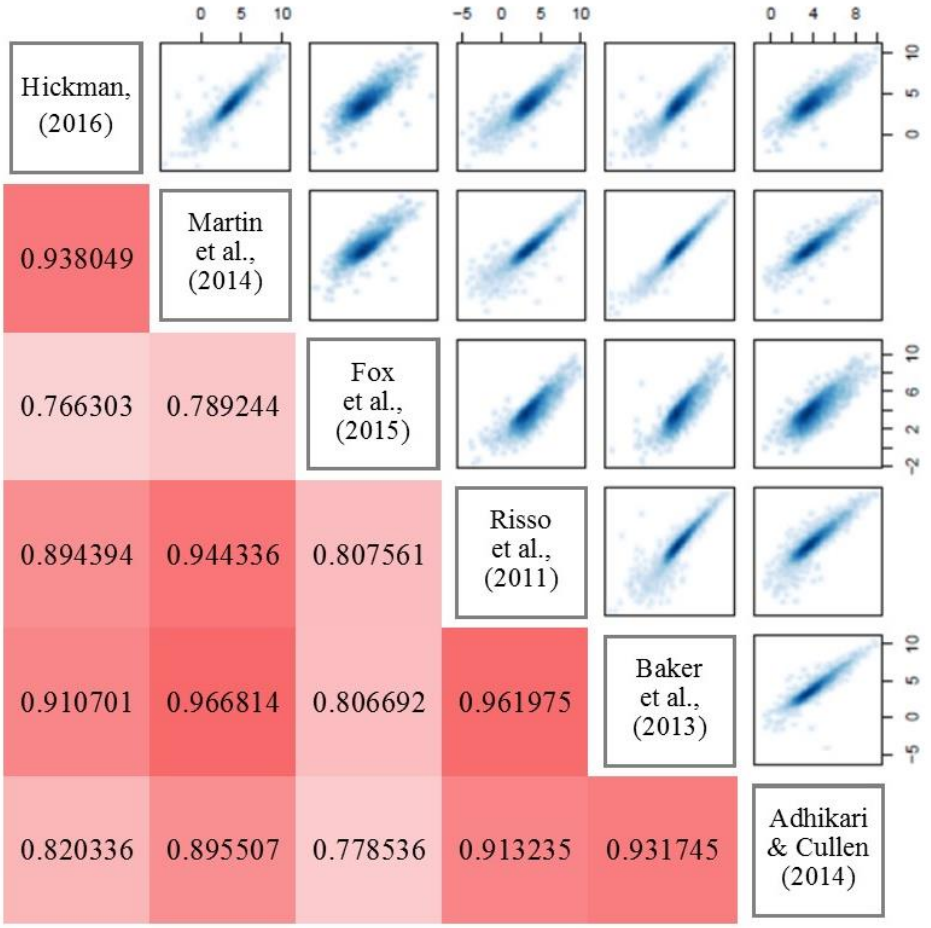

Figure 4. Spearman correlation coefficients and plots for studies in the low variability RPKM dataset

We wanted to test whether the expression rank of each gene (1=lowest expressing gene, 2=next lowest, etc) also does not vary, as opposed to the absolute expression level. Thus, we created rank RNA-seq datasets as described in the Methods section. The correlations for the RNA-seq data normalized by the rank tended to be slightly lower than those normalized by RPKM. In the rank dataset, the Spearman correlations coefficients ranged from 0.688 to 0.924 (see Figure 5). Because the Spearman correlations coefficients did not increase overall from the rank dataset to the low variability rank dataset, the low

variability dataset was not used for further analysis (see Figure 6). The p values of these both of Spearman's rank correlations were $< 2.2\ e^{-16}$.

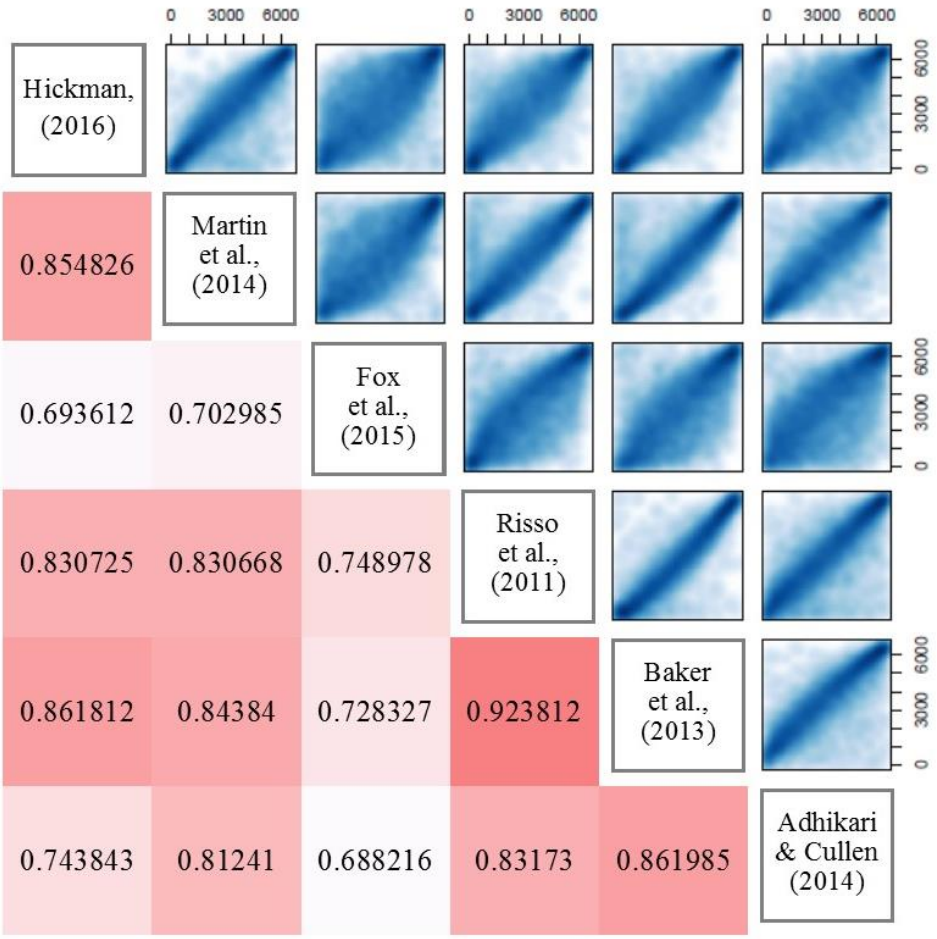

Figure 5. Spearman correlation coefficients and plots for studies in the Rank dataset

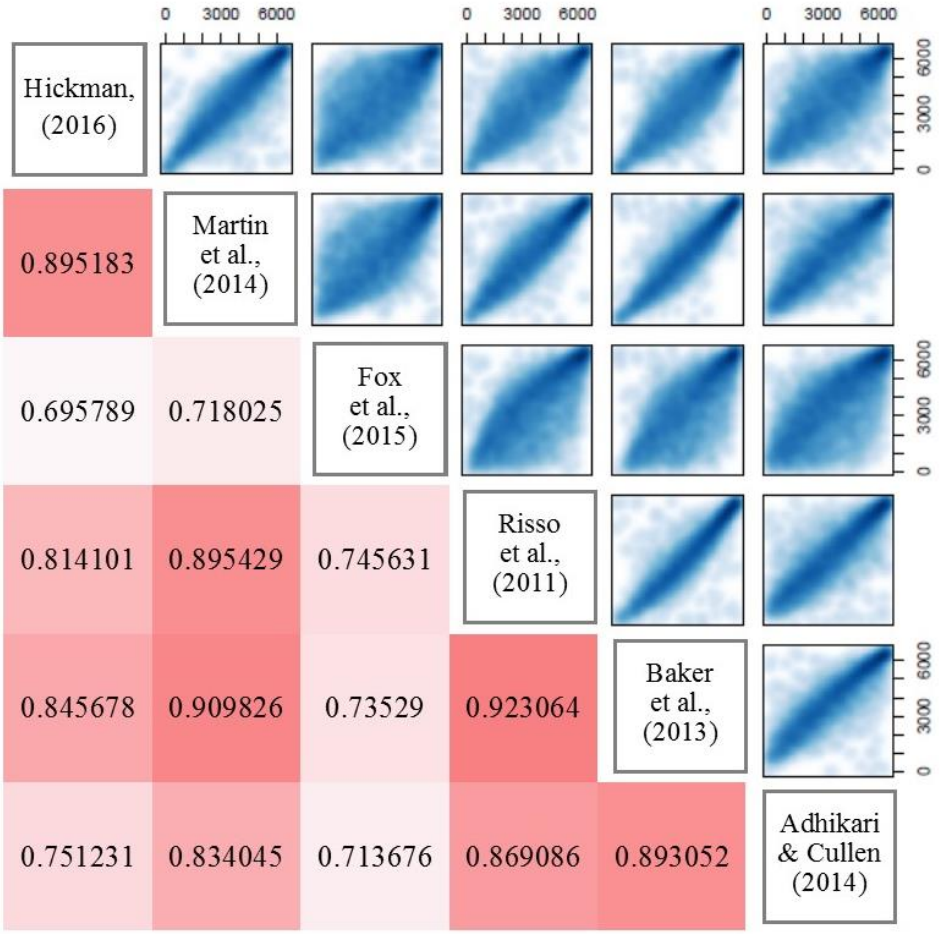

Figure 6. Spearman correlation coefficients and plots for studies in the low variability rank dataset

Correlations and Plots of Protein Datasets

The Spearman correlation coefficients for protein datasets were somewhat lower than those for the normalized RNA-seq datasets. These correlation coefficients ranged from 0.605 to 0.875 (see Figure 7).

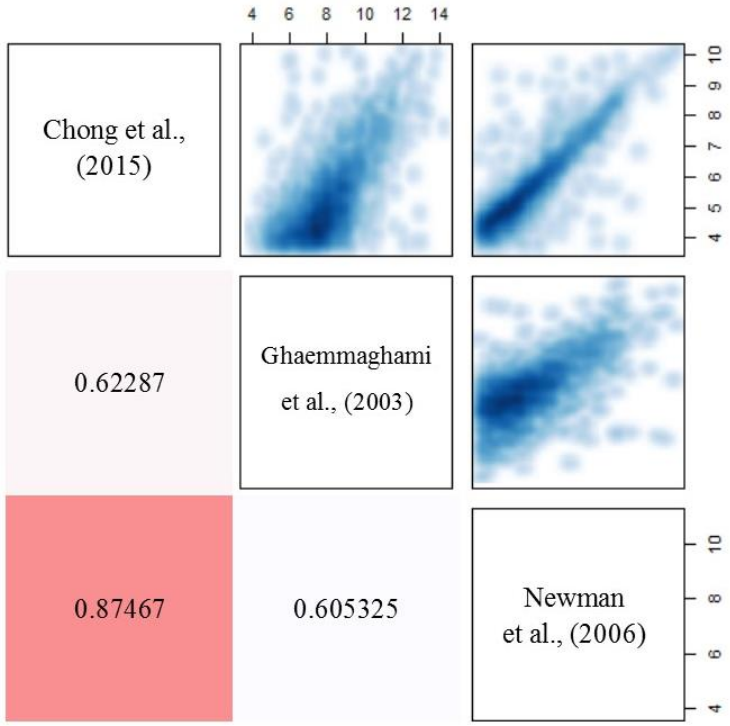

Figure 7. Spearman correlation coefficients and plots for protein abundance datasets.

Correlations and Plots of Normalized RNA-seq Datasets with Other Data

Because RNA-seq values were used as a proxy for protein expression in some portions of this study, it was important to correlate the RNA-seq datasets with protein abundance datasets to confirm an acceptably linear relationship. We plotted the RNA-seq values against the protein abundance values obtained by Chong et al. (2015), Newman et

al. (2006) and Ghaemmaghami et al. (2003). The log plots of each protein abundance study against average RPKM show linear patterns (Figure 8). Consequently, the Spearman correlations coefficients of the protein abundance values in each study compared to the RPKM values were relatively high, ranging from 0.62 to 0.73 (Table 2*)*.

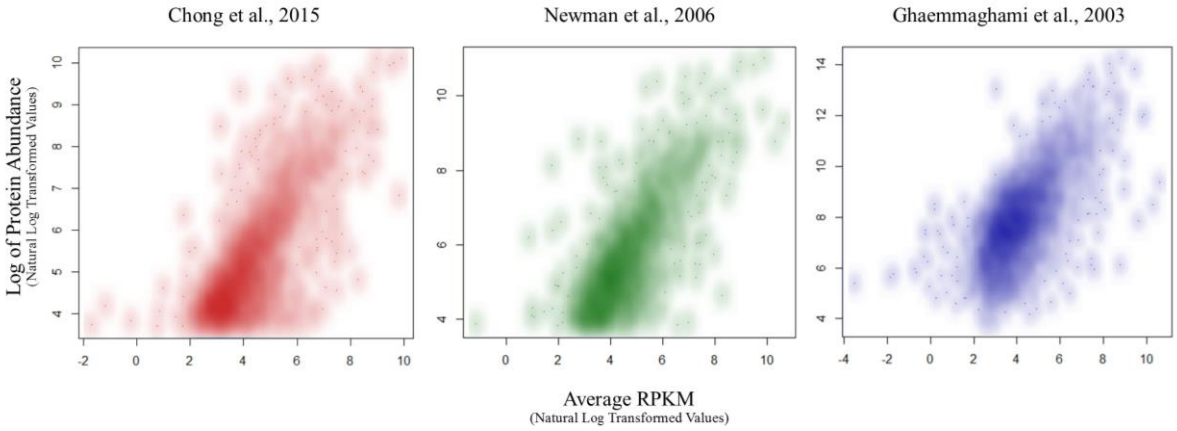

Figure 8. The natural log of average RPKM values (with low variability) was plotted against the natural log of the protein abundances values obtained by Chong et al. (2015), Newman et al.,(2006), and Ghaemmaghami et al. (2003).

Protein abundance datasets were normalized by rank for comparison with the average rank RNA-seq dataset. The resulting plots are shown in Figure 9. The Spearman correlation coefficients of protein abundance ranks compared to RNA-seq ranks, which ranged from 0.65 to 0.70, are shown in Table 2.

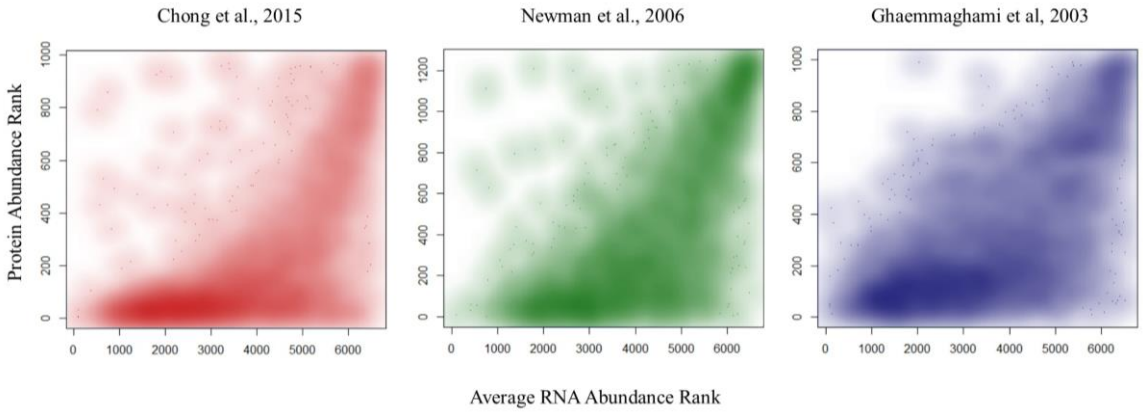

Figure 9. The ranks of RNA-seq values were plotted against ranks of protein abundances derived from the datasets by Chong et al. *(2015),* Newman et al. (2006), and Ghaemmaghami et al. (2003).

In the past, researchers have obtained RNA abundance measurements through other methods, such as Affymetrix microarray. As a comparison, we have plotted the averages of our RNA-seq data from two normalized datasets against Affymetrix microarray data obtained by Legronne et al. (2004). The resulting plots are shown in Figure 10. The Spearman correlation coefficients indicated a high amount of similarity between RNA-seq and Affymetrix microarray values and are shown in Table 2. However, the lowest expressed genes (bottom 10% of expressed genes in Affymetrix microarray dataset) showed much lower correlations (RPKM: r_s= 0.337, Rank: r_s= 0.179). The genes with highest expression (top 10% of expressed genes in the Affymetrix microarray dataset) also showed lower correlations (RPKM: r_s=0.567, Rank: r_s= 0.543).

Table 2

Rna-seq Spearman's Rank Correlation Coefficients

	Study	Spearman Correlation Coefficient (r_s)	
		Average RPKM vs. Protein Abundance	RNA Abundance Rank vs Protein Abundance Rank
Protein Abundance	Chong et al. (2015)	0.6839492	0.6572164
	Newman et al. (2006)	0.7342717	0.706059
	Ghaemmaghami et al. (2003)	0.6244115	0.6129629
RNA abundance by Affymetrix Microarray	Study	Spearman correlation coefficient	
		Affymetrix Microarray Values vs. RPKM Values	Affymetrix Microarray Rank vs. RNA-seq Rank
	Lengronne et al. (2004)	0.8378342	0.8236243

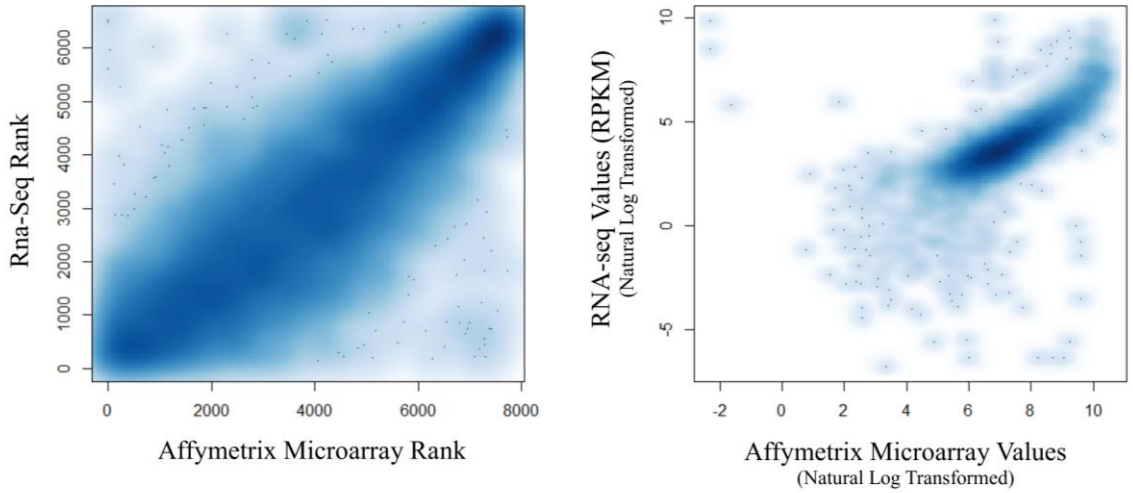

Figure 10. This figure shows the plots of normalized RNA-seq data against RNA abundance data normalized by rank *(left)* or transformed by natural log *(right)* obtained through affymetrix microarray by Legronne et al. (2004).

Descriptive Categories Sorted by Mean

The next question that we wanted to answer was whether RNA or protein expression level was related to protein function or localization. To test this, the mean expression values for each localization and GO category were compared for the RPKM, low variability RPKM, and protein abundance (Newman et al., 2006) datasets. Table 3 shows the localization category means sorted from lowest to highest by the RPKM dataset values, and Table 4 shows the GO category means sorted from lowest to highest by the RPKM dataset values. The values are color coded, with white indicating the lowest mean in the sample and red indicating the highest mean in the sample.

Table 3

Comparison of RNA-seq and Protein Dataset Means for each Localization Category

	Category Average		
Localization Category	RPKM Dataset	Low Variability RPKM Dataset	Protein Abundance Dataset (Newman et al., 2006)
microtubule	27.9	26.8	184.0
spindle_pole	30.8	30.6	117.0
endosome	42.1	43.7	143.4
bud_neck	46.4	45.6	203.2
peroxisome	56.4	45.8	362.7
mitochondrion	83.3	86.3	294.0
bud	84.7	80.4	385.6
nuclear_periphery	87.4	86.9	390.2
actin	92.5	97.6	485.5
late_golgi	97.7	100.2	372.4
lipid_particle	103.1	106.7	651.0
golgi	120.3	121.4	426.8
punctate_composite	121.1	132.1	589.5
er_to_golgi	125.8	125.8	604.7
early_golgi	126.7	126.2	472.7
nucleus	142.9	153.4	826.0
vacuolar_membrane	147.8	148.2	703.1
cell_periphery	154.8	168.6	655.2
nucleolus	183.4	185.8	503.5
vacuole	246.9	285.8	639.4
er	296.5	300.8	650.0
cytoplasm	307.9	292.1	1505.9

Table 4

Comparison of RNA-seq and Protein Dataset Means for each GO Category

GO Category	Category Average		
	RPKM Dataset	Low Variability RPKM Dataset	Protein Abundance Dataset (Newman et al., 2006)
DNA recombination	39.2	40.4	178.8
peroxisome organization	45.8	47.9	291.5
microtubule organizing center	47.0	50.3	346.2
nucleic acid binding transcription factor activity	53.3	47.1	180.8
histone binding	54.3	50.1	283.9
transposition	55.0	52.7	280.3
chromosome segregation	55.9	60.4	325.1
cell morphogenesis	56.0	69.0	399.0
DNA replication	56.9	55.5	552.5
peroxisome	57.2	62.6	265.9
organelle fission	58.7	62.7	424.2
tRNA processing	59.3	59.6	232.5
regulation of DNA metabolic process	59.3	56.0	427.1
protein lipidation	59.9	60.7	161.8
transcription factor binding	63.3	68.2	256.8
mitochondrial translation	64.4	67.9	273.6
helicase activity	65.8	69.7	698.3
endosomal transport	65.9	62.8	214.8
oligosaccharide metabolic process	66.2	66.0	531.2
transcription factor activity, protein binding	67.8	75.4	266.1
chromatin binding	67.9	68.2	266.7
mitotic cell cycle	68.9	71.6	355.6
vitamin metabolic process	71.3	88.2	261.1
protein dephosphorylation	72.9	76.1	304.9
protein binding, bridging	73.5	76.1	402.2
DNA-templated transcription, elongation	75.1	74.1	365.4
cytoskeleton	77.0	83.7	371.0
transcription from RNA polymerase II promoter	77.3	80.5	341.7
regulation of transport	78.9	82.8	250.3
mRNA processing	79.6	84.0	370.2
methyltransferase activity	79.9	81.4	529.4
nuclease activity	79.9	86.6	188.6
DNA repair	80.5	90.1	480.5
RNA splicing	80.9	85.8	303.5
protein acylation	81.6	88.4	298.4
chromosome	81.6	90.5	460.6
DNA-templated transcription, termination	82.0	84.6	581.6
phosphatase activity	82.1	87.3	297.7
regulation of cell cycle	82.2	90.7	544.7

Table 4 (continued)

lipid transport	83.5	74.2	469.6
snoRNA processing	84.4	91.1	257.1
ubiquitin-like protein binding	84.8	92.7	410.4
cell cortex	85.0	90.1	356.2
enzyme binding	86.2	92.2	367.9
cytokinesis	88.0	93.2	281.2
meiotic cell cycle	88.2	100.8	659.0
amino acid transport	89.3	100.0	889.4
protein modification by small protein conjugation or removal	89.6	96.2	1012.3
cytoskeleton organization	89.7	94.0	336.4
peptidyl-amino acid modification	92.7	102.9	501.0
response to osmotic stress	93.3	96.7	947.9
pseudohyphal growth	93.8	97.7	600.3
cytoskeletal protein binding	95.3	99.6	332.7
vesicle organization	95.9	91.9	288.5
membrane invagination	96.4	104.8	509.8
cellular response to DNA damage stimulus	97.1	109.0	589.5
telomere organization	98.4	103.2	953.8
nucleus	100.0	104.0	631.8
RNA modification	100.7	102.8	303.9
transcription from RNA polymerase I promoter	100.7	101.6	390.8
mitochondrion organization	100.9	94.5	608.3
DNA binding	101.2	108.9	570.6
nucleus organization	102.1	111.1	460.3
peptidase activity	103.4	106.7	351.1
proteolysis involved in cellular protein catabolic process	105.2	111.2	611.0
protein maturation	105.8	110.2	378.6
lipid binding	106.9	114.4	394.8
DNA-templated transcription, initiation	107.6	110.1	350.6
protein complex biogenesis	109.3	110.7	454.9
Golgi apparatus	110.5	113.8	351.6
kinase activity	113.2	123.8	1152.6
signaling	113.4	107.4	416.8
transcription from RNA polymerase III promoter	115.4	111.2	407.1
exocytosis	116.4	126.3	256.0
histone modification	117.2	131.8	679.7
hydrolase activity	117.7	125.4	776.3
nucleolus	118.8	126.0	518.6
lipid metabolic process	119.7	117.2	510.6
protein phosphorylation	120.9	138.2	569.1
response to heat	123.1	133.0	423.9
chromatin organization	123.5	132.3	537.5
transferase activity	123.9	133.6	788.8
Golgi vesicle transport	124.6	128.6	434.9
cell budding	125.0	138.2	401.1
endomembrane system	125.9	125.8	432.9
protein targeting	126.8	132.7	914.1
protein alkylation	129.5	145.0	624.2

Table 4 (continued)

carbohydrate transport	129.9	185.9	2416.8
endocytosis	130.9	140.9	479.7
cellular respiration	133.7	143.3	391.9
cytoplasmic, membrane-bounded vesicle	134.7	136.7	576.4
nucleotidyltransferase activity	136.5	143.9	1168.7
sporulation	138.2	166.0	1023.2
ion binding	142.5	126.8	1672.7
response to starvation	144.2	160.1	738.1
RNA catabolic process	145.9	161.4	574.4
transferase activity, transferring glycosyl groups	147.1	164.4	336.2
invasive growth in response to glucose limitation	148.2	177.9	192.4
transmembrane transporter activity	148.3	159.2	760.4
hydrolase activity, acting on glycosyl bonds	148.4	170.6	184.6
transmembrane transport	149.5	156.8	1288.9
regulation of organelle organization	150.3	166.9	447.5
cellular ion homeostasis	150.9	169.7	697.9
mitochondrial envelope	154.4	138.7	567.2
endoplasmic reticulum	156.2	148.5	497.3
guanyl-nucleotide exchange factor activity	156.6	144.5	545.0
protein transporter activity	158.3	156.8	636.4
ion transport	158.7	172.3	833.8
ATPase activity	161.1	176.2	1397.8
GTPase activity	161.6	161.7	583.0
site of polarized growth	167.3	177.5	451.2
ligase activity	172.7	176.2	1464.6
membrane	174.6	184.2	932.1
regulation of protein modification process	181.2	201.8	1031.2
protein glycosylation	183.7	188.8	983.7
cellular bud	187.6	201.3	447.4
response to chemical	189.7	201.9	713.9
enzyme regulator activity	192.6	220.3	841.2
organelle inheritance	196.4	176.0	449.2
mRNA binding	204.1	232.5	1000.1
vacuole	207.0	221.3	1311.5
mitochondrion	213.3	220.2	1299.4
conjugation	229.3	280.9	708.7
tRNA aminoacylation for protein translation	230.5	230.5	2053.1
signal transducer activity	232.5	340.1	304.6
cytoplasm	250.9	245.8	1079.5
cellular amino acid metabolic process	273.0	307.4	1683.7
response to oxidative stress	281.6	301.2	1182.0
translational initiation	285.8	262.1	1715.0
protein folding	302.4	336.6	2485.5
organelle fusion	308.5	335.4	467.7
cell wall organization or biogenesis	324.0	365.1	1143.3
oxidoreductase activity	327.2	376.5	1885.4
unfolded protein binding	342.7	363.3	3352.2
membrane fusion	344.3	377.9	493.5
ribosomal subunit export from nucleus	347.5	380.0	991.9

Table 4 (continued)

RNA binding	356.3	376.6	2053.0
plasma membrane	360.1	417.2	2121.7
vacuole organization	375.9	421.9	513.7
extracellular region	379.9	412.7	543.8
rRNA processing	421.2	479.7	1312.2
nuclear transport	431.2	456.3	2093.3
isomerase activity	480.5	401.5	1287.3
ribosomal large subunit biogenesis	487.7	538.6	1645.2
regulation of translation	491.2	517.0	1350.9
translation factor activity, RNA binding	504.2	385.1	5294.7
nucleobase-containing compound transport	508.5	535.4	2472.5
carbohydrate metabolic process	525.9	635.2	3052.6
organelle assembly	552.6	599.5	1385.5
lyase activity	583.9	676.6	3606.5
ribosomal small subunit biogenesis	593.8	676.9	1398.6
rRNA binding	642.2	788.8	1203.5
nucleobase-containing small molecule metabolic process	706.7	844.9	3767.1
cell wall	788.1	1034.3	4548.3
cofactor metabolic process	789.0	913.8	4258.2
generation of precursor metabolites and energy	799.7	1045.8	4162.3
monocarboxylic acid metabolic process	900.3	1086.0	6099.1
ribosome	1162.6	1015.6	2553.2
ribosome assembly	1164.2	1298.7	2062.1
structural molecule activity	1185.4	1057.4	2204.0
translational elongation	1414.7	1418.4	6144.7
structural constituent of ribosome	1645.7	1407.8	2826.3
cytoplasmic translation	2145.8	1823.5	3914.4

Bootstrapping

Localization bootstrapping. To determine which localization categories contain genes with statistically high or low expression, a bootstrapping statistical analysis compared expressions levels in each category to a set of randomly chosen genes that were not in the category. The localization categories that contain statistically significant values are shown in Figures 11-14. Specifically, for this analysis, we used the RPKM dataset (Figure 11), low variability RPKM dataset (Figure 12), Rank dataset (Figure 13), and protein abundance dataset (Figure 14). The categories with statistical significance are similar between each analysis. All four analyses found that cytoplasm genes had statistically significant high expression levels, and spindle pole and mitochondrion genes had statistically significant low expression levels.

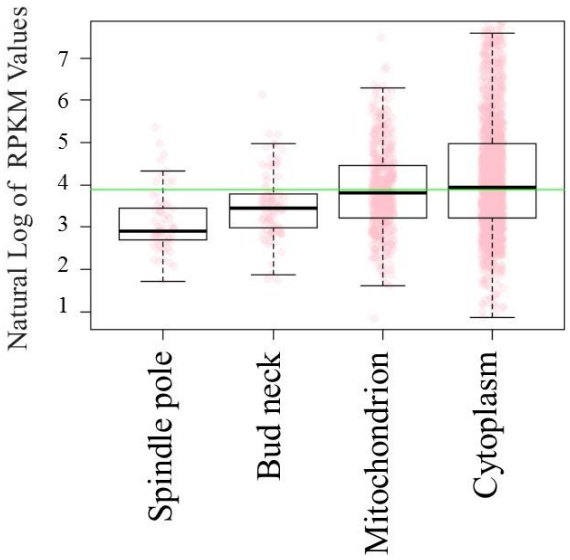

Figure 11. Statistically Significant Localization Categories in the RPKM Dataset. Boxplots are shown for localization categories with p values less than 0.005. Pink dots represent individual RPKM values and the green horizontal line represents the average of all localization category means.

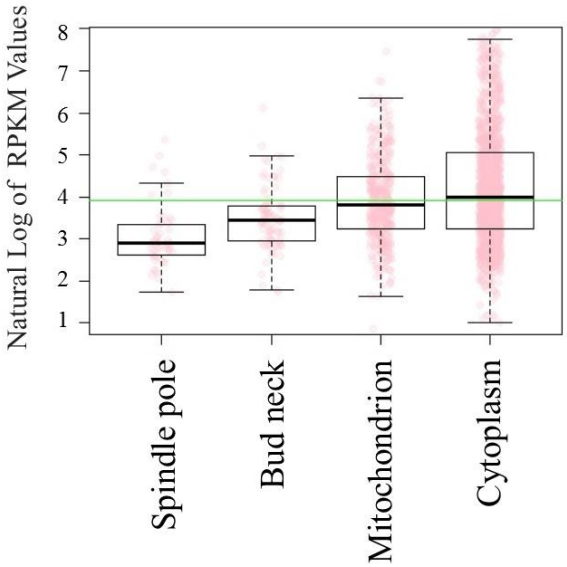

Figure 12. Statistically Significant Localization Categories in the Low Variability RPKM Dataset. Boxplots are shown for localization categories with p values less than 0.005. Pink dots represent individual low variability RPKM values and the green horizontal line represents the average of all localization category means.

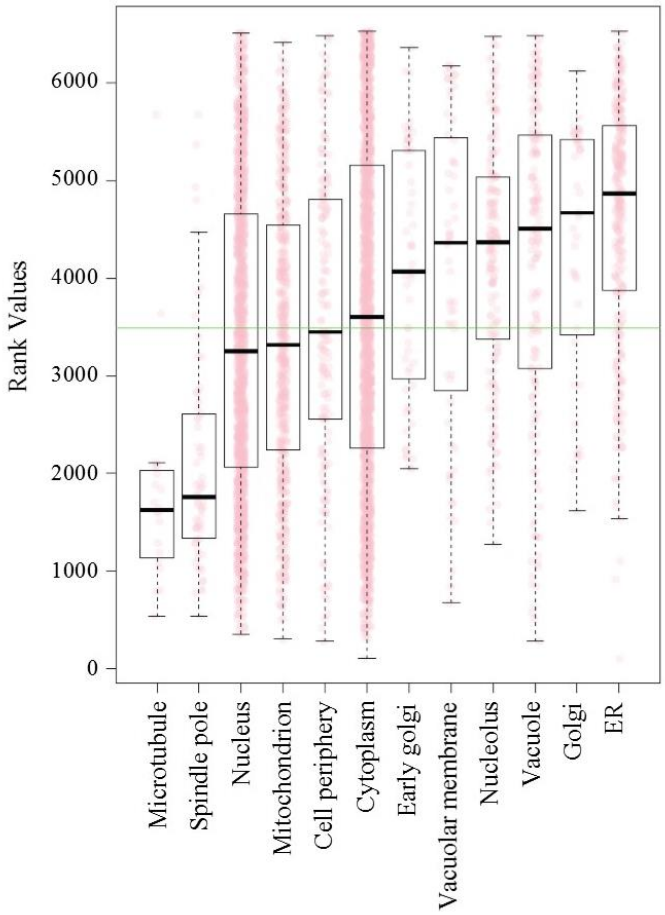

Figure 13. Statistically Significant Localization Categories in the Rank Dataset. Boxplots are shown for localization categories with p values less than 0.005. Pink dots represent individual rank values and the green horizontal line represents the average of all localization category means.

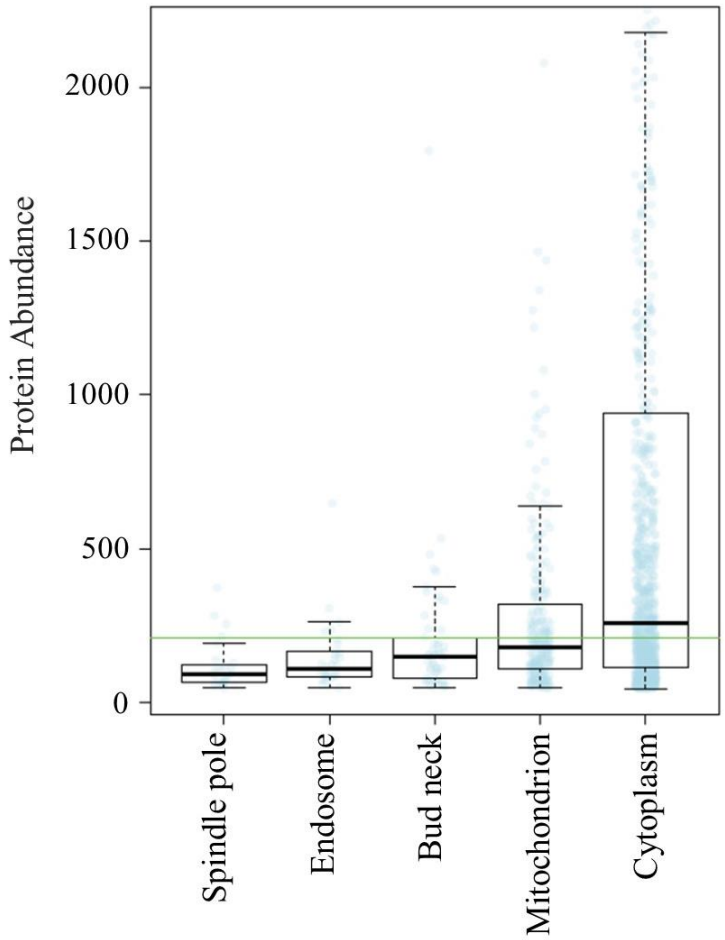

Figure 14. Statistically Significant Localization Categories in the Protein Dataset. Boxplots are shown for localization categories with p values less than 0.005. Blue dots represent individual protein abundance values and the green horizontal line represents the average of all localization category means.

GO bootstrapping. To determine which GO categories contained genes with statistically significant high or low expression values, as above, the bootstrapping statistical analyses compared the expression levels in each GO category to a random set of genes. The statistically significant localization categories for RPKM are shown in Figure 15, for low variability RPKM are shown in Figure 16, for rank are shown in Figure 17, and for protein abundance are shown in Figure 18. The categories with

statistical significance are similar between each analysis. All four analyses found that cytoplasmic translation, structural constituent of the ribosome, ribosome, nucleobase containing small molecule metabolic process, structural molecule activity, cofactor metabolic process, and cytoplasm had statistically significant high distributions and nucleus, transcription from RNA polymerase II promotor, nucleic acid binding transcription factor activity, and DNA recombination had statistically significant low distributions.

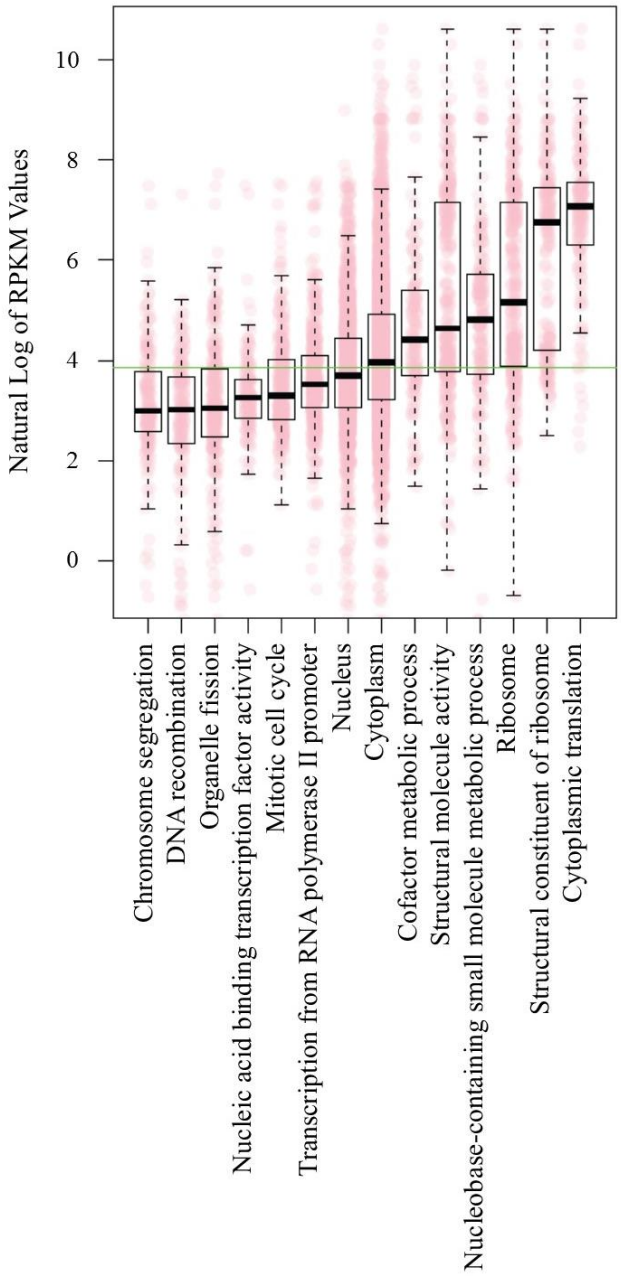

Figure 15. Statistically Significant GO Categories in the RPKM Dataset. Boxplots are shown for GO categories with p values less than 0.003. Pink dots represent individual RPKM values and the green horizontal line represents the average of all GO category means.

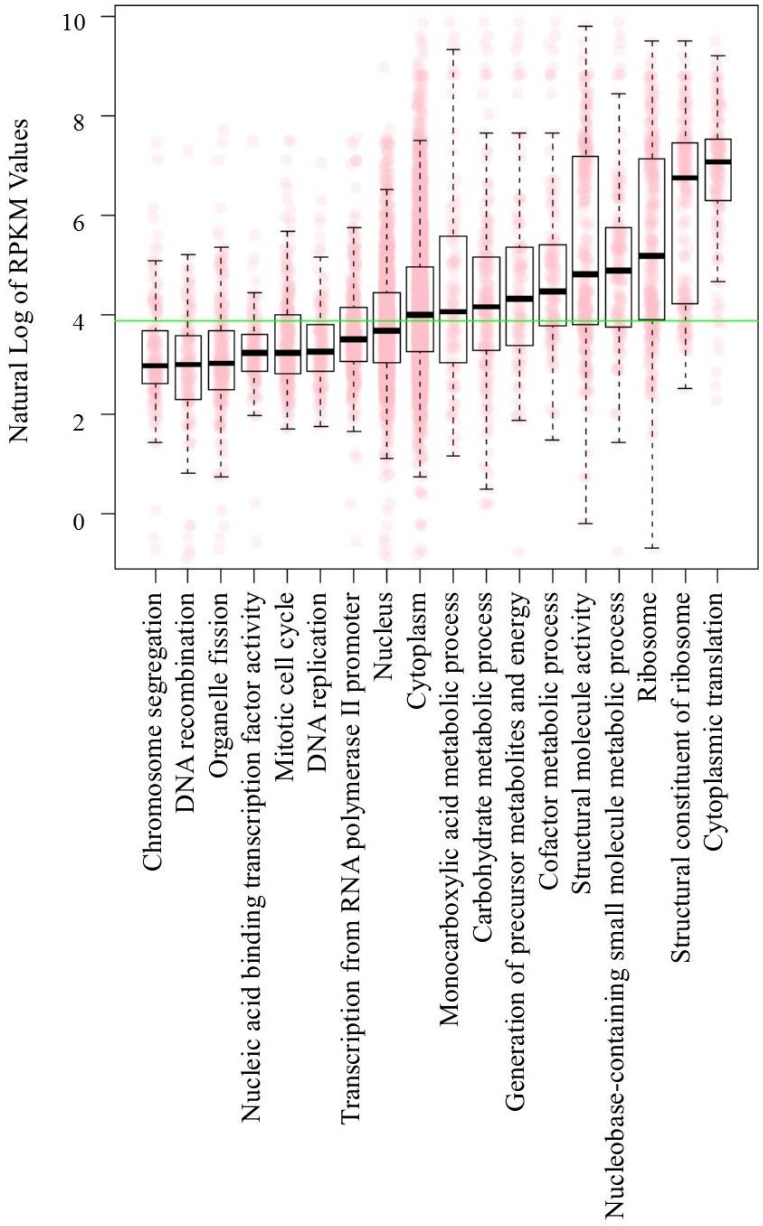

Figure 16. Statistically Significant GO Categories in the Low Variability RPKM Dataset. Boxplots are shown for GO categories with p values less than 0.003. Pink dots represent individual RPKM values and the green horizontal line represents the average of all GO category means.

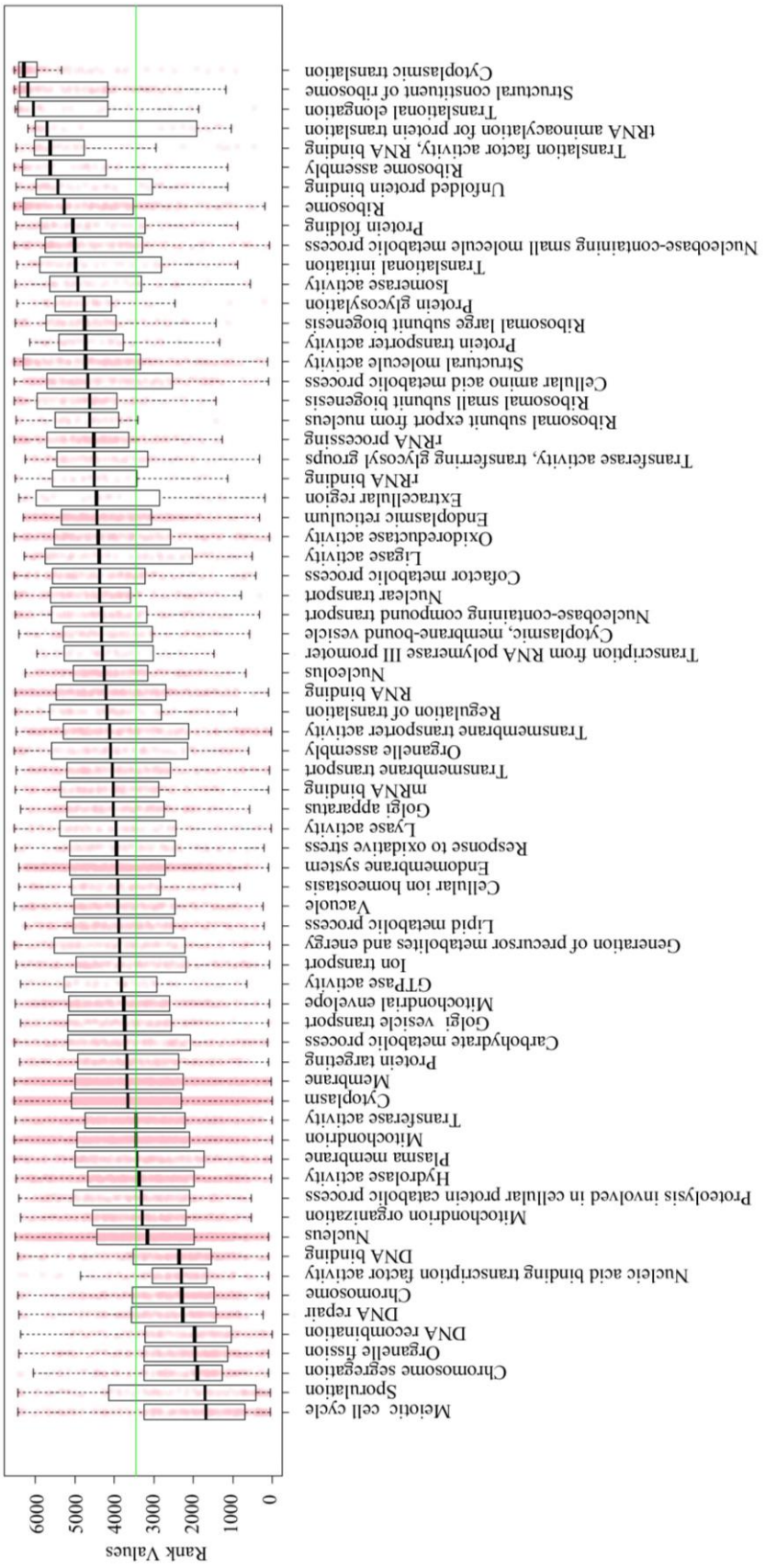

Figure 17. Statistically Significant GO Categories in the Rank Dataset. Boxplots are shown for GO categories with p values less than 0.003. Pink dots represent individual rank values and the green horizontal line represents the average of all GO category means.

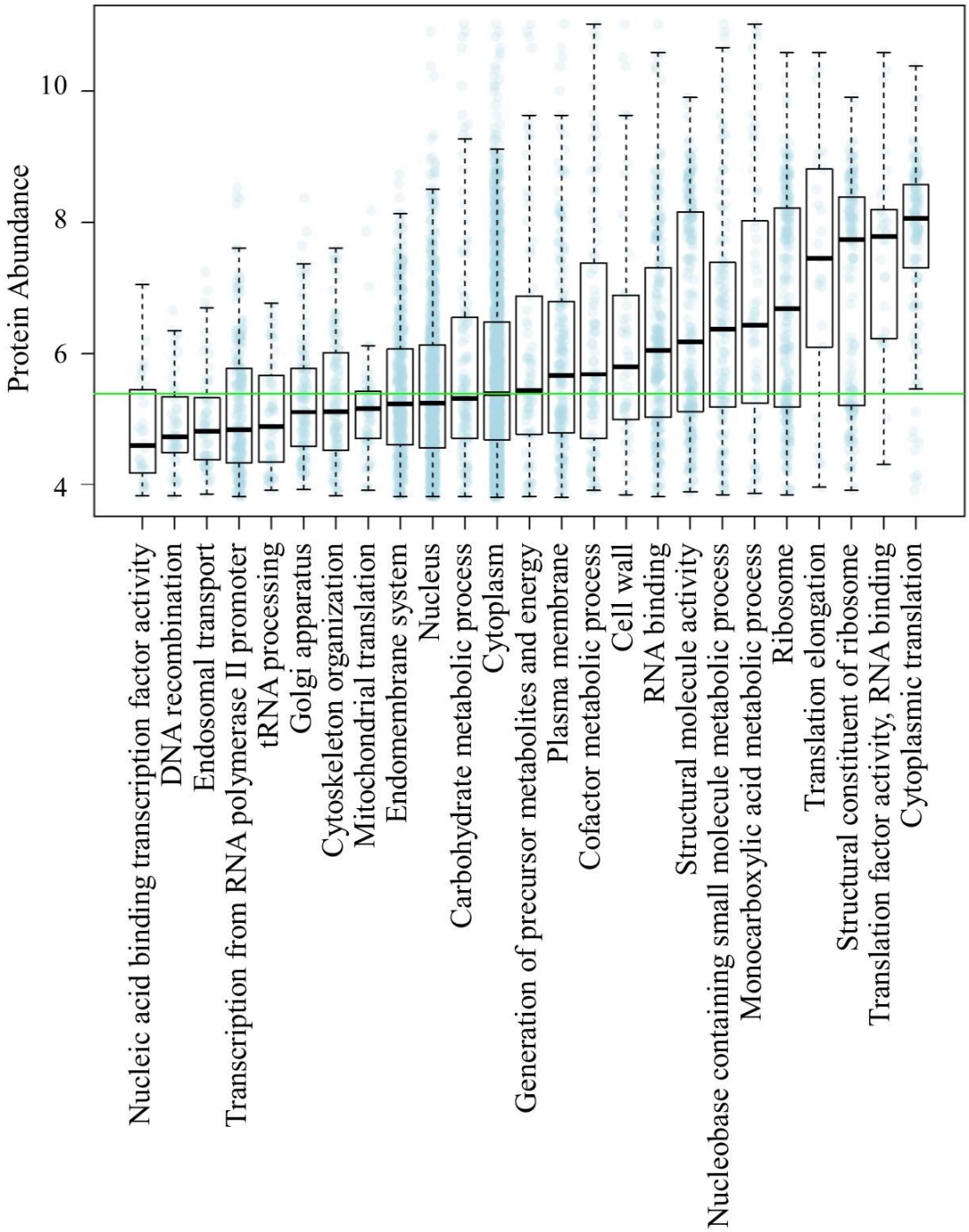

Figure 18. Statistically Significant GO Categories in the Protein Abundance Dataset. Boxplots are shown for GO categories with p values less than 0.003. Blue dots represent individual protein values and the green horizontal line represents the average of all GO category means.

Artificial Neural Network Results

Normalized RNA-seq data. In order to determine whether the GO and localization descriptions could predict gene expression levels, we used an artificial neural network approach. Gene descriptions (i.e., GO annotations or localization) were used in the input layer and expression levels (i.e., RPKM, RPKM Low Variability, Rank) were used in the output layer throughout the training and testing of the neural network. The neural network's ability to predict gene expression levels throughout 50 repetitions is shown in Figure 19. GO categories were able to predict binary RNA-seq levels with an average of 75.1 % accuracy. Localization categories were able to predict binary RNA-seq levels with an average of 66 % accuracy. As a control, GO and localization values were only able to predict the random binary vector with an average of 50.6 % accuracy, as expected. Individualized average prediction accuracy percentages are shown in Table 5.

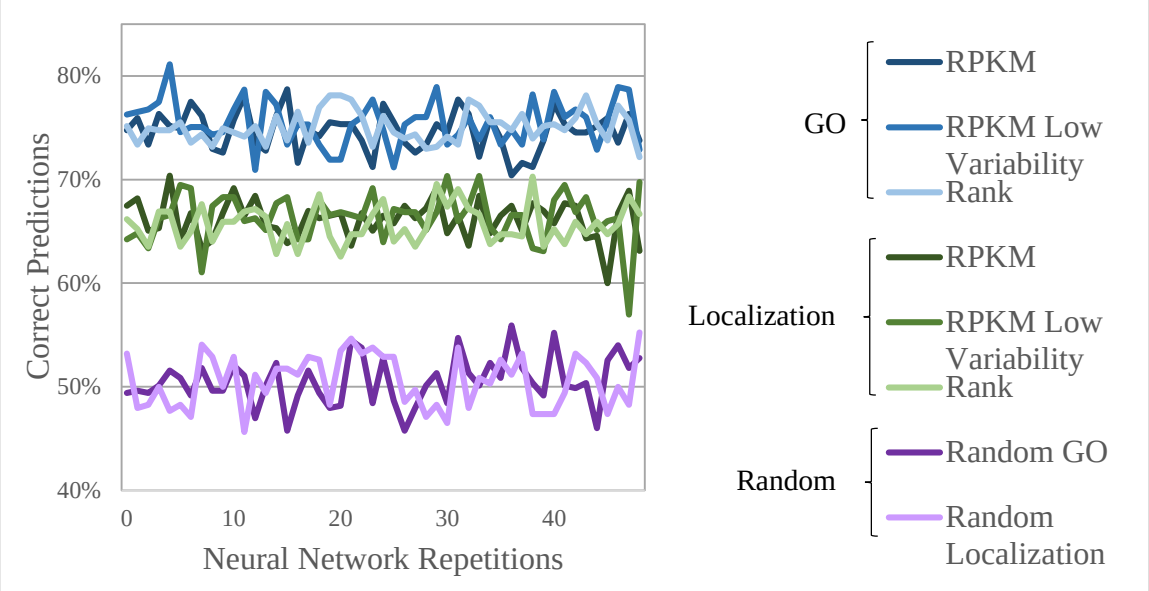

Figure 19. Prediction Accuracy of Artificial Neural Network for RNA-seq data. The prediction accuracy is shown for 50 repetitions of the neural network for all dataset combinations.

Table 5

Average Prediction Accuracy of Artificial Neural Network for Normalized RNA-seq Datasets

Combinations of datasets used for input variables (left) and output variables (right)		Average Percentage of Correct Predictions
Go	RPKM	74.7%
	Low Variability RPKM	75.6%
	Rank	75.1%
	Random	50.5%
Localization	RPKM	66.1%
	Low Variability RPKM	66.4%
	Rank	65.6%
	Random	50.6%

Protein abundance data. The ability of localization or GO categories to predict protein expression level was tested as above using the neural network. Gene descriptions (i.e., GO annotations or localization) were used in the input layer and protein expression levels, obtained from the Newman et al., (2006) protein abundance dataset, were used in the output layer throughout the training and testing of the neural network. The neural network's ability to predict gene expression levels throughout 50 repetitions is shown in Figure 20. GO categories were able to predict binary protein abundance levels with an average 69.9% accuracy. Localization categories were able to predict binary protein abundance levels with an average of 66.2% accuracy. In contrast, GO and localization values were only able to predict the random binary vector with an average of 50.65 % accuracy, as expected. Individualized average prediction accuracy percentages are shown in Table 6.

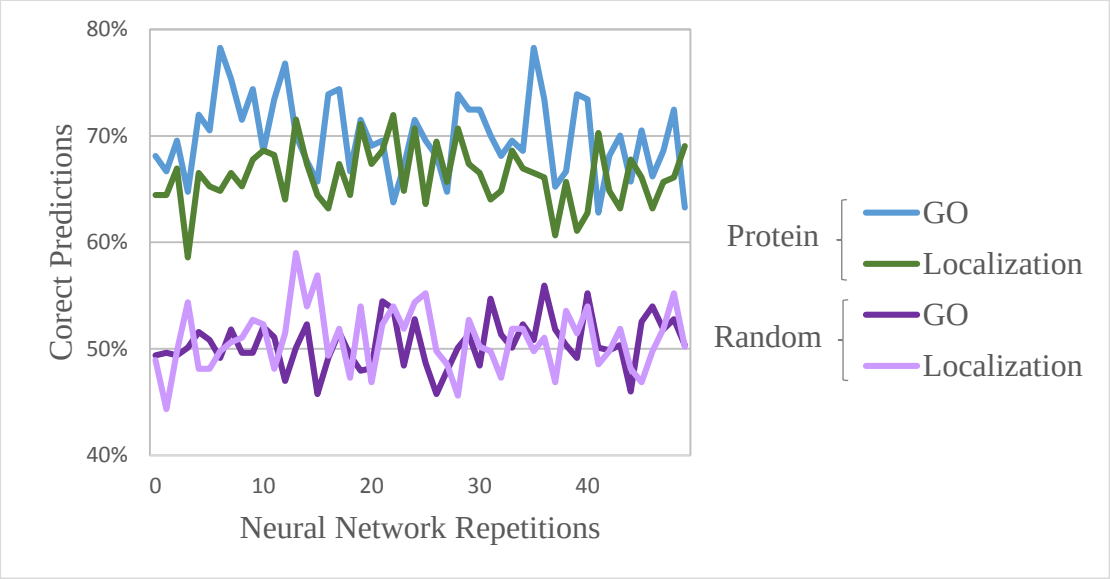

Figure 20. Prediction Accuracy of Artificial Neural Network for Newman et al. (2006) Protein Abundance data. The prediction accuracy is shown for 50 repetitions of the neural network for all dataset combinations.

Table 6

Average Prediction Accuracy of Artificial Neural Network for Newman et al. (2006) Protein Abundance Dataset

Combinations of datasets used for input variables (left) and output variables (right)		Average Percentage of Correct Predictions
Go	Protein	69.9%
	Random	50.5%
Localization	Protein	66.2%
	Random	50.8%

Chapter 4

Discussion

Correlation Analysis

RNA-seq is an effective method for consistently quantifying genome-wide expression levels. Here, we demonstrated the consistency of steady-state expression levels measured by RNA-seq quantifications by correlating *S. cerevisae* expression data obtained in similar conditions by six different research teams. The similarity in expression levels was high, as genome-wide correlations coefficients ranged from 0.684 to 0.964. This high amount of consistency suggests the expression level of each gene is fixed. The slightly lower correlations of Fox et al. (2015) to all other samples may be a result of the platform used, which may vary slightly in sensitivity (SEQC/MAQC-III Consortium, 2014). While we observed high reproducibility in the RNA-seq expression levels measured in *S. cerevisiae* under steady-state conditions, the reproducibility of RNA-seq measurements has been previously reported by other researchers who have examined expression levels in different cell types, including cells from mouse tissues (Mortazavi et al., 2008), and human tissues (Marioni, et al., 2008).

Previously, researchers have found good correlations between Affymetrix microarray data and RNA-seq data when quantifying differential expression in *S. cerevisiae* (Nookaew et al., 2012). However, the correlations have not been shown to be as strong for genes expressed at low levels (Mortazavi, et al., 2008). In our study of steady-state gene expression, we also observed a high degree of similarity between RNA-seq data and Affymetrix microarray data, with correlation coefficients ranging from 0.823 and 0.837. In agreement with previous studies, we found that genes with the lowest

expression had much lower correlations, and genes with high expression showed somewhat lower correlations. This finding gives further support for the utilization of the RNA-seq method for the study of genome-wide expression.

Although protein abundance sets were gathered from varying strains, conditions, and methods, the correlation coefficients of protein abundance datasets compared to each other were fairly high ranging from 0.605 to 0.875. In comparison to the RNA-seq datasets which spanned the majority of the transcriptome (5552-7130 genes represented in each set), the protein abundance datasets contained a much smaller collection of data (2,385-3868 genes represented in each set). While the protein and RNA-seq data sets were not perfectly correlated with each other, their correlation coefficients, ranging from 0.613 to 0.734, showed that there is a high degree of similarity between protein abundance and RNA-seq expression levels. In previous research, the correlations between mRNA and protein abundance data in yeast have widely varied with Pearson correlation coefficients (R^2) ranging from 0.34 to 0.87 (Abreu, et al., 2009). Our data may have shown higher correlations in comparison to some of the correlations previously obtained by researchers due to our use of data obtained through the very consistent RNA-seq method, rather than previously used methods such as microarray.

Limitations in RNA-seq & Protein Abundance Measurements

The differences observed between the RNA-seq count levels and protein abundance levels could be due to a variety of factors. Post-transcriptional factors likely attribute to some of this difference (Guimaraes, Rocha, & Arkin, 2014). The differences could also be due to the size of the datasets used in this study. As mentioned earlier, the protein datasets that were collected measured the protein abundance coded by

approximately 2385 to 3868 genes, while the RNA-seq datasets spanned more of the genome, covering between 5552 and 7130 genes. Thirdly, while variability was reduced through normalization methods, some differences could be due to the variations in methods used by different researchers. The importance of using both mRNA and protein expression data was confirmed through the comparison of category means and bootstrapping.

Expression Trends

Similarities between RNA-sequencing and protein abundance methods were found when mRNA and protein datasets were merged with localization and GO categories. As shown in Tables 3 and 4, categories with low expression averages in the RNA-seq datasets tended to have low expression averages in the protein datasets. Additionally, categories with high expression averages in the RNA-seq datasets tended to have high expression averages in the protein datasets as well. In general, genes with GO categories relating to transcription and DNA-related processes tended to have lower expression averages, while genes relating to the ribosome and translation tended to have higher expression averages. This trend was confirmed by the bootstrapping results.

GO categories that were shown to contain genes with statistically significant high expression levels in all expression datasets (3 normalized RNA-seq datasets and 1 protein abundance dataset) were cytoplasmic translation, structural constituent of the ribosome, ribosome, nucleobase containing small molecule metabolic process, structural molecule activity, cofactor metabolic process, and cytoplasm. GO categories with statistically significant low expression levels, shown in all expression datasets, were nucleus,

transcription from RNA polymerase II promoter, nucleic acid binding transcription factor activity, and DNA recombination (see Figures 11- 18).

In the bootstrapping analysis of expression levels in localization categories, cytoplasm was shown to contain statistically significant highly expressed genes in all four bootstrapping analyses (3 normalized RNA-seq datasets and 1 protein dataset). The categories mitochondrion and spindle pole were shown to have statistically significant low distributions in all four bootstrapping analyses. Our results are similar to those found by Drawid, Jansen, and Gerstein (2000) who compared the gene expression levels of proteins localized to eight subcellular localization compartments in yeast cells. In their study, the cytoplasm was shown to have high expression levels and mitochondrion was shown to have one of the lowest expression levels. Expression levels of the nucleus tended to be low in our study, but did not always show statistically significant low expression. However, in the study by Drawid, Jansen, and Gerstein (2000), the nucleus was shown to have the lowest expression levels. This difference may be due to the additional localization categories that were introduced in our study.

The results of the artificial neural network repetitions confirmed that there are biological processes, molecular functions, cellular components, and localizations that differ in the genes that are expressed at high levels versus genes that are expressed at low levels, such that the neural network was able to predict binary mRNA or protein levels using binary GO or localization categories with 66.1 to 75.6 percent accuracy. The prediction accuracy was more similar for mRNA and protein levels predicted by localization categories with the average mRNA prediction accuracy of 66.0% and the average protein prediction accuracy of 66.2%. The mRNA levels predicted by GO

categories were higher for mRNA levels (average of 75.1%) than protein levels (69.9%). This variation in prediction accuracy may be due to the smaller number of genes in the protein dataset compared to the large number of genes in the normalized RNA-seq datasets. If a smaller number of genes is used, the patterns may not be as easily detectable. In all cases, the prediction accuracy was higher for mRNA and protein values than random vector values. This confirms that the portion of the predictions exceeding ~50% is based on the presence of true patterns in the data, and not just a result of random chance.

Conclusion

In this study, we've demonstrated the reproducibility of the RNA-seq method for the quantification of mRNA levels in *S. cerevisiae* under consistent conditions. This high level of reproducibility indicates that the level of each gene is fixed under constant conditions.

Both the mRNA and protein abundance levels tended to be high for genes involved in translation and the ribosome, and low for genes involved in transcription and DNA-related processes. This may indicate that the *S. cerevisiae* needs high levels of translation and ribosomal proteins to survive, such that the high cost required to produce these proteins is outweighed. On the other hand, the yeast cell may not need a large amount of protein to complete the transcription and DNA-related processes necessary for the survival of the cell. These levels can remain low to balance the energy costs of the cell. This may suggest the genes expressed at low levels are non-essential, are needed in only low concentrations, or function in a smaller volume within the cell.

Statistically significant high expression levels were shown for cytoplasmic translation, structural constituent of the ribosome, the ribosome, nucleobase containing small molecule metabolic process, structural molecule activity, cofactor metabolic process, and the cytoplasm. Statistically significant low expression levels were shown for the nucleus, transcription from RNA polymerase II promoter, DNA recombination, nucleic acid binding transcription factor activity, the spindle pole, and the mitochondrion. These categories represent potential evolutionary pressures for the high or low expression of genes in yeast.

The distinction in gene ontology and localization characteristics of genes which were highly expressed, compared to those which were expressed at low levels, was confirmed through the use of an artificial neural network.

Future Research

At the current time, the gene ontology or localization categories of some genes in *S. cerevisiae* are identified as ambiguous or unknown. As the categorizations of more genes are identified, it will be important to reanalyze the expression levels in each category, and to rerun the neural network analysis to obtain more accurate results.

This research should also be carried out in other organisms using a combination of RNA-seq and protein abundance data. The expression levels of the genes of the same cell type in other organisms should be measured under standard conditions to confirm that the level of a gene's expression is fixed under consistent conditions. Our findings would be further supported if the GO and localization categories of other organisms show similar expression trends to those shown in our study.

References

Abreu, R. de S., Penalva, L.O., Marcotte, E.M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular bioSystems, 5*(12), 1512-1526. doi:10.1039/b908315d.

Adhikari, H., & Cullen, P.J. (2014). Metabolic respiration induces AMPK- and Ire1p-dependent activation of the p38-Type HOG MAPK pathway. *PLoS Genet.* 10 (10). doi: 10.1371/journal.pgen.1004734.

Albert, F.W., Treusch, S., Shockley, A.H., Bloom, J.S., & Kruglyak, L. (2014). Genetics of single-cell protein abundance variation in large yeast populations. *Nature, 506(7489),* 494-497. doi:10.1038/nature12904.

Baker, L.A., Ueberheide, B.M., Dewell, S., Chait, B.T., Zheng, D., & Allis, C.D. (2013). The yeast Snt2 protein coordinates the transcriptional response to hydrogen peroxide-mediated oxidative stress. *Mol Cell Biol. 33* (19): 3735-3748. doi: 10.1128/MCB.00025-13.

Booth, F.W., & Lees, S.J. (2007). Fundamental questions about genes, inactivity, and chronic diseases. *Physiological Genomics, 28* (2): 146-157.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P., & Boeke, J.D. (1998). Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmid for PCR-mediated gene disruption and other applications. *Yeast, 14:* 115-132.

Brawand, D., Soumillon, M., Nesculea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., & Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature, 478*(7369), 343-348.

Cooper, G.M. (2000). *The Cell: A Molecular Approach. 2nd edition*. Southerland, MA: Sinauer Associates Inc.

Chong, Y.T., Koh, J.L.Y., Frieson, H., Duffy, S.K., Cox, M.J., Moses, A., Moffat, J., Boone, C., & Andrews, B.J. (2015). Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell, 161,* 1413-1424.

Erdman, S., Lin, L., Malczynski, M., & Snyder, M. (1998). Pheromone-regulated genes required for yeast mating differentiation. *The Rockefeller University Press, 140* (3), 461-483. doi: 10.1083/jcb.140.3.461

Fox, M.J., Gao, H., Smith-Kinnaman, W.R., Liu, Y., Mosley, A.L. (2015). The exome component Rrp6 is required for RNA polymerase II termination at specific targets of the Nrd1-Nab3 pathway. *PLoS Genet. 11(2*). doi: 10.1371/journal.pgen.1004999.

Fu, X., Fu,N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., Khaitovich, P. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics, 10*(161). doi: 10.1186/1471-2164-10-161

The Gene Ontology Consortium. (2005). Documentation. Retrieved from http://geneontology.org/page/documentation

Ghaemmaghami, S., Huh, W., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S. (2003). Global analysis of protein expression in yeast. *Nature, 425(16),* 737-741.

Goldberg, A.L. (2003). Protein degradation and protection against misfolded or damaged proteins. *Nature, 426,* 895-899.

Guimaraes, J.C., Rocha, M., & Arkin, A.P. (2014). Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli. Nucleic Acids Research, 42(8).*

Günther, F., and Fritsch, S. (2010). Neuralnet: training of neural networks. The R Journal, 2(1), 30-38.

Hackett, N.R., Butler, M.W., Shaykhiev, R., Salit, J., Omberg, L., Rodriguez-Florez, J.L., Mezey, J.G., Strulovici-Barel, Y., Wang, G., Didon, L., & Crystal, R.G. (2012). RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics, 13*(82). doi: 10.1186/1471-2164-13-82

Hannan, R.D., Jenkins, A., Jenkins, A.K., & Bradenburger, Y. (2003). Brief review: Cardiac Hypertrophy: A matter of translation. *Clinical and Experimental Pharmacology and Physiology, 30,* 517-527.

Hu, Z., Chen, K., Xia, Z., Chavez, M., Pal, S., Seol, J.H., Chen, C.C., Li, W., & Tyler, J.K. (2014). Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. doi: 10.1101/gad.233221.113.

Huh, W., Falco, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O-Shea, E.K. (2003). Global analysis of protein localization in budding yeast. *Nature, 425,* 686-691.

Jackowiak, P., Nowacka, M., Strozycki, P.M., & Figlerowicz, M. (2011). RNA degradome- its biogenesis and functions. *Nucleic Acids Research,* 1-10. doi: 10.1093/nar/gkr450

Jenner, L., Melnikov, S., de Loubresse, N.G., Ben-Shem, A., Iskakova, M., Urzhumtsev, A., Meskaushas, A., Dinman, J., Yusupova, G., & Yusupov, M. (2012). Crystal Structure of the 80S yeast ribosome. *Current Opinion in Structural Biology, 22*(6), 759-767. doi: 10.1016/j.sbi.2012.07.013

Krishnamurthy, S., Hampsey, M. (2009). Eukaryotic transcription initiation. *Current Biology, 19* (4), R153-R156. doi:10.1016/j.cub.2008.11.052

Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., Kelly, G.P., Itoh, T., Watanabe, Y., Shirahige, K., & Uhlmann, F. (2004). Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature, 430(6999),* 573-580. doi: 10.1038/nature02742

Li, J.J., Bickel, P.J., & Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ, 2:e270.* doi: 10.7717/peerj.270

Li, Y., Zhang, W., Zheng, D., Zhou, Z., Yu, Wenwen, Zhang, L., Liang, X., Guan, W., Zhou, J., Chen, J., & Lin, Z. (2014). Genomic evolution of *Saccharomyces cerevisiae* under Chinese rice wine fermentation. *Genome Biol Evol. 6(9),* 2516-2526. doi: 10.1093/gbe/evu201.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research,* 18:1509-1517. doi: 10.1101/gr.079558.108

Martin, G.M., King, D.A., Green, E.M., Garcia-Nieto, P.E., Alexander, R., Krogan, N.J., Gozani, O.P., & Morrizon, A.J. (2014). Set5 and Set1 cooperate to repress gene expression at telomeres and retrotransposons. *Epigenetics. 9(4),*513-522. doi: 10.4161/epi.27645.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., & Wold, B. (2008a). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*(7), 621-628.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., & Wold, B. (2008b). Mapping and quantifying mammalian transcriptomes by RNA-Seq, Supplementary Figures and Text. *Nature Methods, 5*, 621-628.

Nagalakshmi, U., Wang, Z., Waern, K., Schou, C., Raham D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science, 320*(5881), 1344-1349.

Newman, J.R.S., Ghammaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., & Weissman, J.S. (2006). Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature, 441,* 840-846. doi:10.1038/nature04785

Nookaew, I., Papini, M., Pornputtpong, N., Scalcinati, G., Fagerberg, L., Uhlen, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisae*. *Nucleic Acids Research,* 1-14. doi:10.1093/nar/gks804

Nuzhdin, S.V., Wayne, M.L., Harmon, K.L., & McIntyre, L.M. (2004). Common pattern of evolution of gene expression level and protein sequence in *Drosophila. Mol. Biol. Evol., 21*(7), 1308-1307. doi: 10.1093/molbev/msh128

Parker, R. (2012). RNA degredation in *Saccharomyces cerevisiae. Genetics, 191,* 671-702.

Patro, S.G.K., & Kumar sahu, K. (n.d.). Normalization: A preprocessing stage. Retrieved from http://arxiv.org/ftp/arxiv/papers/1503/1503.06462.pdf

Pestova, T.V., Kolupaeva, V.G., Lomakin, I.B., Pilipenko, E.V., Shatsky, I.N., Agol, V.I., & Hellen, C.U.T. (2001). Molecular mechanisms of translation initiation in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America, 98*(13), 7029-7036. http://doi.org/10.1073/pnas.111145798

Ponnala, L., Wang, Y., Sun, Q., & van Wijk, K.J. (2014). Correlation of mRNA and protein abundance in developing maize leaf. *The Plant Journal, 78:* 424-440. doi: 10.1111/tpj.12482

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Rifkin, S.A., Kim, J., White, K.P. (2003). Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genetics, 33,* 138-144. doi:10.1038/ng1086

Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinformatics. 12(480).* doi: 10.1186/1471-2105-12-480.

Rung, J., & Brazma, A. (2012). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics Advanced Online Publication,* 1-11. doi:10.1038/nrg3394

Ruscio, J. (2008). Constructing confidence intervals for Spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. Journal of Modern Applied Statistical Methods, 7(2), 416-434.

Schwanhäusser, B., Busse1, D., Li1, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, W. (2011). Global quantification of mammalian gene expression control. Nature, 473, 337-342. doi:10.1038/nature10098

SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology, 32(9), 903-914. doi:10.1038/nbt.2957

SGD Project. (2015, April 8). SGD features. Retrieved from http://downloads.yeastgenome.org/curation/chromosomal_feature/SGD_features.README

SGD Project. (2016a, February 14). Go Slim Mapping. Retrieved from http://downloads.yeastgenome.org/curation/literature/go_slim_mapping.tab

SGD Project. (2016b, April 15). Saccharomyces cerevisiae Genome Snapshot. Retrieved from http://www.yeastgenome.org/genomesnapshot

Stoebel, D.M., Dean, A.M., & Dykhuizen, D.E. (2008). The cost of expression in Escherichia coli lac operon proteins is in the process, not in the products. Genetics, 178, 1653-1660. doi: 10.1534/genetics.107.085399

Wagner, A. (2005). Energy constraints on the evolution of gene expression. Molecular Biology and Evolution, 22(6), 1365-1374.

Wagner, A. (2007). Energy costs constrain the evolution of gene expression. Journal of Experimental Zoology (Mol Dev Evol), 308B, 322-324.

Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D., and Brown, P.O. (2002). Precision and functional specificity in mRNA decay. PNAS, 99(9), 5860-5865. doi /10.1073/pnas.092538799

Wang, Z., Gerstein, M., & Snyder, M., (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10, 57-63. doi:10.1038/nrg2484

Yu, J., Xiao, J., Ren, X., Lao, K., & Xie, S. (2006). Probing Gene expression in live cells, one protein molecule at a time. Science, 311(5767), 1600-1603. doi: 10.1126/science.111962