USING PATHWAY CORRELATION PROFILES FOR UNDERSTANDING PATHWAY PERTURBATION

A Dissertation

presented to

the Faculty of the Graduate School

University of Missouri – Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

ALLISON TEGGE

Dr. Dong Xu, Dissertation Supervisor

DECEMBER 2012

The undersigned, appointed by the dean of the Graduate School, have examined the

dissertation entitled

USING PATHWAY CORRELATION PROFILES FOR UNDERSTANDING PATHWAY PERTURBATION

presented by Allison Tegge,

a candidate for the degree of

Doctor of Philosophy

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Dong Xu

Dr. Chi-Ren Shyu

Dr. Jianlin Cheng

Dr. Mark Hannink

DEDICATION

To mumsie and papaya.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Dong Xu, for his continued support, encouragement, and guidance as I complete my studies. Second, I would like to thank my committee for their comments, suggestions, and time through this whole process. Lastly, I would like to thank everybody else that has been here for me throughout this time, encouraged and distracted me, and answered all my questions and asked me even more.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

USING PATHWAY CORRELATION PROFILES FOR UNDERSTANDING PATHWAY PERTURBATION

ALLISON TEGGE

Dr. Dong Xu, Dissertation Supervisor

ABSTRACT

Identifying perturbed or dysregulated pathways is critical to understanding the biological processes that change within an experiment. Previous methods identified important pathways that are significantly enriched among differentially expressed genes; however, these methods cannot account for small, coordinated changes in gene expression that amass across a whole pathway. In order to overcome this limitation, we developed a novel computational approach to identify pathway perturbation based on pathway correlation profiles. In this approach, we can rank the pathways based on the significance of their dysregulation considering al gene-gene pairs. We have shown this successfully for differences between two experimental conditions in *Escherichia coli* and changes within time series data in *Saccharomyces cerevisiae*, as well as two estrogen receptor response classes of breast cancer. Overall, our method made significant predictions as to the pathway perturbations that are involved in the experimental conditions.

Further, I can use these pathway correlation profiles to better understand pathway dynamics and modules of regulation. I have applied this developed method to the Ribosome pathway for several model organisms and various tissue types, where I was able to isolate alternative regulation patterns for each species and tissue. In addition, I have applied these pathway correlation profiles for the MAPK pathway to help

characterize the disease progression of colon cancer from normal tissue, through all four stages, culminating in final metastasis. The pathway correlation profile method allows for more meaningful and biologically significant interpretation of the current data available.

In short, we developed a novel computational method for identifying pathway perturbation. This method is a powerful tool that better utilizes gene expression data when studying pathway dynamics in regards to biological processes. Moreover, this method provides hypotheses for understanding the mechanisms within meaningful pathways, and where the pathway dynamics change across conditions.

1 Chapter 1: Introduction

Understanding biological systems as a whole offers a global perspective, but obtaining this view is not trivial, and in many instances not always possible. The current experimental data being generated typically provide a single snapshot of a biological situation, either experimental condition or disease state, and in many instances this data is analyzed on a single entity level, such as individual gene or reaction. Though there has been an increase in efforts to try to incorporate multiple data types, most previously generated data sets are not structured or organized for the ease of such meta-analyses. In many cases, previously developed data sets only assay one data type (e.g., gene expression, protein level, etc.) for that specific experiment, and not multiple data types. To compensate, more computational method need to be developed that help bridge the gap between single modality data sets and a systems biology perspective, while using preexisting publically available data sets.

1.1 High-throughput experiments

With the current technologies available, it is no longer necessary to interrogate a small set of genes. Instead, high-throughput experiments that assay the entire genome are commonplace. With this increase in large data sets, there is a high demand for computational methods to be able to handle, process and analyze the data. Currently, there are numerous relatively straightforward computational methods for handling the high-throughput data, e.g., identifying mutations, and differential gene or protein expression. Alternatively, and arguably more important, there are few complex computational approaches that additionally consider a second level of information, such

as gene or protein interactions, or multiple data types. These more complex approaches is where an emphasis needs to be made for developing computational models, which take into account a systems-wide view of the biological processes. In this dissertation, we try to develop a computational method that sets the foundation for these higher-level analyses of high throughput genomic data.

1.2 Gene expression

Gene expression profiling analysis has successfully been used in identifying causal genes for many diseases. However, there are many cases where several genes work together in a synergistic fashion, thus resulting in a complex disease. For these situations, standard gene expression analysis approaches are not as effective and robust at distinguishing the causal genes, thereby missing several conclusions from the data sets. In examples like these, more specialized approaches need to be utilized that examine multi-gene effects. Such multi-gene effects could be the result tandem differential expression of two genes, or even an unstructured expression profile between several genes. For whichever processes of multi-gene effects are employed, computational methods need to be developed to handle these additional conclusions. We try to develop an advanced approach that utilizes gene expression data, but does not rely on absolute expression levels from any particular gene. In our approach, we can aim to identify those genes that work in either a structured or unstructured manner, resulting in a specific phenotype.

1.3 Publically available data sets

Nowadays, experiments are being designed with a meta-analysis in mind. However, with the current plethora of data sets available, especially those that are in public repositories, it should not be compulsory to run more experiments to gain a new insight on the current

research foci and questions. Instead, an interest should be made towards developing computational methods that utilize the publically available data in new and novel ways. By using this approach, we can gain new perspectives and new knowledge can be investigated and learned that was not previously addressed from the data under other analytical methods. Moreover, if the publically available data is reanalyzed, especially when asking new and more relevant questions, future experiments could be prioritized for a more efficient and cost effective future analysis.

1.4 Dissertation organization

This dissertation is organized into several parts. Chapter 2 introduces several of the key concepts that are the foundation to the pathway correlation profiles. Chapter 3 presents the methodology and benchmarking for the novel pathway correlation profile method for identifying pathway perturbation. Chapter 4 and 5 discuss applications and the usefulness of correlation profiles, as well as the interpretations that can be made thereof. Chapter 6 will explore the conclusions for the pathway correlation profile method, such as the advantages and inherent limitations. Lastly, the dissertation will end with Chapter 7 and future areas where pathway correlation profiles could successfully be applied.

2 Chapter 2: Literature Review

Identification of perturbed or dysregulated pathways is important for understanding changes in biological processes between two conditions. Microarray technologies are essential for identifying differences in gene expression, but there has been limited in-depth use of microarray data on a pathway level. When it comes to pathways, microarray data is typically used to identify pathways enriched with significantly differentially expressed genes. Ultimately, these studies try to extrapolate activated/repressed pathways, i.e., those pathways that show global increases and decreases in gene expression, respectively [1]. Alternatively, microarray data can also be used via co-expression networks for pathway reconstruction where little to no prior pathway knowledge is applied in the co-expression networks.

2.1 Gene set enrichment analysis methods

In order to identify pathways of interest, various gene set enrichment (GSE) methods are utilized [2,3,4,5]. These methods rank genes by the expression's signal-to-noise ratios [6] or the correlation of expression with the phenotype [2], determine an enrichment score for each gene ontology or pathway, and then select a set of gene ontologies or pathways based on the significance of their enrichment scores. Keller et al. extended a GSE method by utilizing dynamic programming in order to optimize this selection of significant signaling pathways [7]. GSE methods, however, require a set of genes in the gene list or pathway to be differentially expressed with statistical significance; though this requirement is sufficient in many instances, it is not necessary in order for a pathway to

be dysregulated. Furthermore, this condition may not accurately reflect globally perturbed pathways. There are many biological circumstances where a few differentially expressed genes can be identified; yet large pathological differences are observed, such as diagnosis-relapse events [8,9]. In order to help reduce this dependency on differentially expressed genes, Adewale et al. developed a regression analysis to handle pathway data, where they agglomerate a pathway-level test statistic for each individual gene in the pathway [10]. Again, though this returns a pathway level result, it still looks at each gene individually and not at how the genes coordinate with each other within the pathway. Moreover, rich information in microarray data may be underutilized. For example, current computational methods generally aggregate the biological replicates into a mean or median, thus losing added information from the available data.

2.2 Gene co-expression analysis

Microarray data has also been used for gene-gene correlation or the co-expression of genes, which has resulted in novel pathway identification [11,12,13,14]. In particular, methods for pathway identification often rely on strong correlations between two genes. Inversely, those genes that are not co-regulated are assumed not correlated, which sometimes may not be the case. Due to these limitations, Childs et al. developed both a condition dependent and independent approach for establishing functional annotation modules to describe regulatory processes [14]. As an extension of novel pathway identification, Novak and Jain used selective gene co-expression in order to confirm valid pathways [15]. Allocco et al. also showed that there is a relationship between regulation and co-expression [16]. Through these methods, they identified that there is an increase

in gene-gene correlation when a common mechanism of regulation is involved with both genes, e.g., a common transcription factor. Though these methods utilize gene-gene correlation, their goal is to identify pathways or modules that are co-regulated, yet may not interrogate general perturbations of known or unknown gene sets that are not revealed in the form of co-expression.

To aid in identifying perturbed pathways, differential gene-gene co-expression has been implemented for studying changes between different diseases and biological conditions. Lai et al. use gene-gene co-expression to identify genes with similar co-expression patterns to those that are already known to be involved in the biological process of interest [17]. This method does not rely on differential expression of genes, but relies on coordinated gene expression instead; however, it still interrogates the expression data on a gene level and does not look at the global differential gene-gene co-expression of the pathway. Cho et al. used differential co-expression to identify gene sets (e.g., pathways) that have differences in gene expression [18], but again does not look into the changing dynamics within a pathway or gene set. More applicably, however, Freudenberg et al. used differential co-expression coupled with unsupervised learning in order to identify gene sets that are significant under various conditions [19]. While their method does not utilize previously defined gene sets, it does show that given significant gene sets, there is an increase in gene-gene pair correlations. There are several disadvantages to these methods. They assess the behavior of individual genes to summarize the activity of a pathway and/or do not look at the trends of pair-wise interactions between genes within an entire predefined pathway.

2.3 Gene clustering methods

Clustering gene expression data attempts to group together genes that have similar expression profiles [20]. To accomplish this, several approaches can be used from statistical (e.g., k-means, k-medoids) [21], to data mining (e.g., hierarchical clustering, trees) [22,23], to machine learning (e.g., self-organizing maps, support vector machines) [24,25], to name a few. However, in general each of these methods is a 1-dimensional approach where they assess how closely two genes' expression profiles are to each other, and then group them in the same cluster based on this metric.

This 1-D approach is beneficial to identifying similar behaviors in gene expression between genes, but in many experiments, there are additional variables of interest, for example condition and sample. To include these additional variables and features, gene expression pattern identification has progressed into bi-clustering. Bi-clustering is an approach that not only clusters the genes based on gene expression profiles across conditions, but also clusters conditions based on gene expression profiles across genes [26,27]. These approaches are successful in identifying specific gene expression patterns under specific conditions, however they only consider one gene at a time. This single gene approach forces each gene into one specific cluster. In several cases, constraining a gene into one cluster is sufficient. Biology, however, is not binary; for example, one gene might be responsible for the crosstalk between two pathways [28] or a protein could be involved in multiple complexes [29]. Situations such as these require alternate clustering schemes that will allow a gene in more than one cluster.

Other clustering methods have been applied to gene expression, especially for temporal experiments. Ernst and colleagues developed a method for clustering time series gene

expression [30,31]. In this method, they identify sets of genes with common expression patterns by correlating genes with model profiles. Several additional studies have achieved temporal gene expression clustering by using Bayesian methods [32], hidden Markov models [33] and episode mining [34,35].

All of these methods, however, still consider each gene individually. Though this approach is successful for very distinctive clusters, it is not optimal for identifying groups of genes that have small changes and moreover rarely looks at all of the biological replicates. By considering clustering of gene expression data in an all-together new perspective, where genes are no longer forced into one cluster, we can make predictions that are more biologically relevant and realistic.

2.4 Other methods

To overcome these challenges, we developed a systematic method of using microarray gene expression data to identify pathway perturbation based on changes in pathway correlation profiles derived from the gene-gene pairs. Given gene sets extracted from known pathways, we identify significant pathways based on changes in gene co-expression. We can identify those pathways that are significantly perturbed in an experiment, as well as isolate groups of genes that are known to be strongly involved in a pathway's regulation. In addition, we can identify potential significant genes that may be involved in the perturbation of a pathway but are not differentially expressed as defined by statistical confidence. Our method no longer relies on single gene involvement as well as effectively utilizes the added information gained from biological replicates within an experiment to successfully identify and rank significantly perturbed pathways.

Further, we can use the pathway correlation profiles to investigate regulative dynamics within a pathway. This can be accomplished under a one-state analysis, or multi-state analysis by characterizing the gene-gene relationships under various conditions. Through this, we can identify differential regulation patterns for a pathway under various conditions or time points. This allows us to identify on a more comprehensive level the various means in which the pathway is regulated. This additional feature of the pathway correlation profile method allows extensive and complete analysis of an experiment from gene expression data through pathway identification and interpretation.

3 Chapter 3: Pathway correlation profiles

A general schematic for our pathway correlation profile method is shown in **Figure 1**. Initially, gene expression data is processed and normalized. Expression profiles are then created for the set of genes involved in each pathway. Using these expression profiles, pathway correlation profiles are created for each pathway and pathway perturbation is estimated via bootstrapping. These results are then combined to rank the pathways based on their perturbation. The details for each step are further explained below.

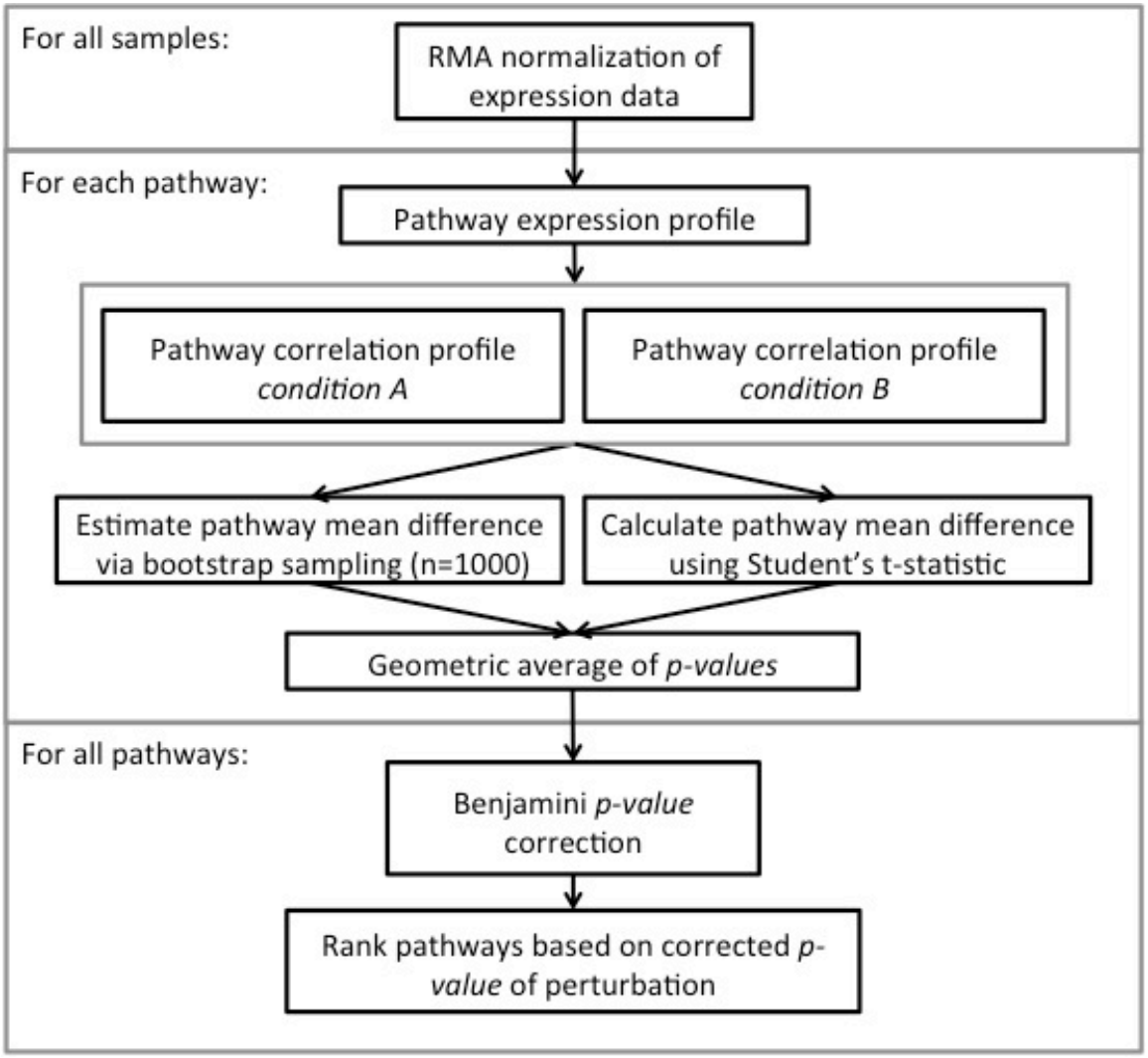


Figure 1. Flowchart describing pathway correlation perturbation method for analyzing gene expression data on a pathway level.

Initially, gene expression data is processed and normalized. Expression profiles are then created for the set of genes involved in each pathway. Using these expression profiles, pathway correlation profiles are created in each condition for each pathway. These results are then combined to determine the pathway's mean difference in gene-gene pair correlations, and then ranked based on their significance of perturbation.

3.1 Methodology

3.1.1 Assumptions

The pathway correlation profiles method for pathway analysis has the assumption that essential genes that are working together are highly correlated; those genes that are non-essential, and subsequently not working together, will show a background correlation profile that is random around zero. To confirm this assumption, we can simulate various levels of perturbation on a pathway. Given a pathway correlation profile where the gene-gene pairs within the pathway show a bias towards positive correlations, we can induce a perturbation by adding a random noise variable to each gene. For each gene's expression, e_i, within the pathway, we can calculate ε_i, an error model for gene i under a known condition.

$$\varepsilon_i = N\left(0, \sigma_i^2 p^2\right)$$

where p is a perturbation and σ is the standard deviation of gene i. We can then add this error parameter to the original expression data to create our simulated expression profiles, e_i^*:

$$e_i^* = e_i + \varepsilon_i$$

By applying the perturbation as a random error proportional to standard deviation of the gene's expression, we can still maintain the structure of gene expression within the pathway by ensuring that those genes with large standard deviations will still have large standard deviations and those with small standard deviations will have small standard deviations, respectively.

Figure 2 shows the effect of simulation different levels of perturbation upon a pathway correlation profile. With small levels of perturbation (d=0.1), we can see that there is minimal change in the pathway correlation profiles. As the levels of perturbation increase (d=0.5, 1, 2), the pathway correlation profiles change drastically, and in fact, migrate towards a distribution centered around zero.

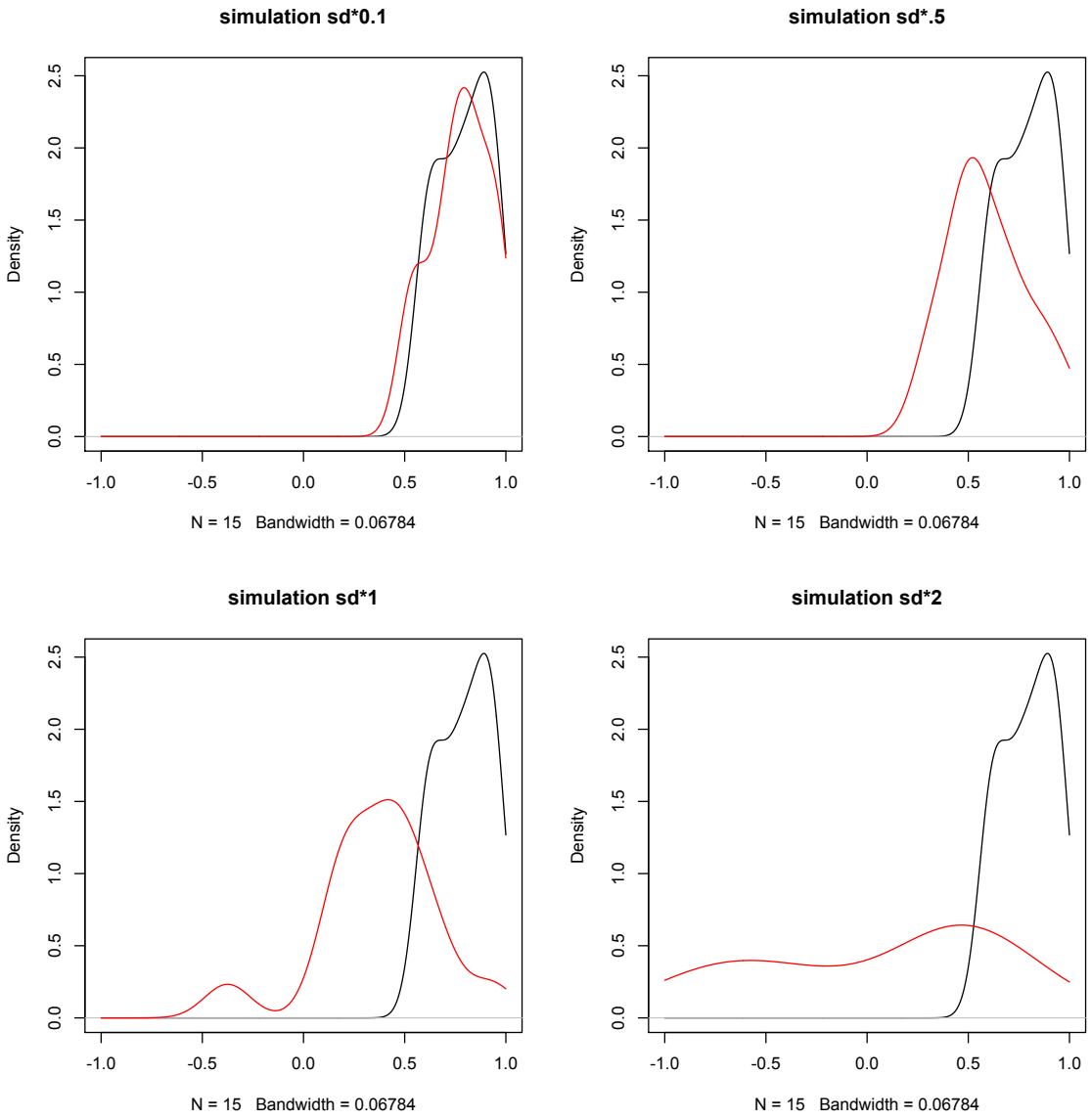


Figure 2. Simulated perturbation of a pathway.

Black: original pathway correlation profile; Red: simulated pathway correlation profile. x-axis: gene-gene pair correlation; y-axis: relative density.

3.1.2 Expression data

E. coli microarray gene expression data were downloaded from the GEO website (GSE4511) [36]. The platform for this data set is the Affymetrix *E. coli* Antisense Genome Array. This data set investigated changes in gene expression when *E. coli* was treated with different pH environments: pH 5.0, pH 7.0, and pH 8.7. In total, there were five separate samples for each pH. In addition, a *S. cerevisiae* time series data set was also downloaded from the GEO website (GSE1311-4) [37]. This data set utilized the Affymetrix Yeast S98 arrays. Singh et al. performed a desiccation in combination with rehydration of *S. cerevisiae* in order to identify transcriptional changes over time. To determine changes over the rehydration process, they performed a time series experiment with nine samples at each of the following time points: 0 (dry), 15, 45, 90 and 360 minutes after rehydration. Samples from a control group were also included. To show the robustness of this method, a breast cancer data set comparing gene expression between positive and negative estrogen receptor (ER-positive and ER-negative) status patients was analyzed [38]. Breast cancer gene expression data set was downloaded from the GEO website (GSE2034) and is from Affymetrix Human U133a GeneChips with 77 ER-negative and 209 ER-positive patient samples.

3.1.3 Pathway data

Pathway data were collected from the KEGG database [39], including the metabolic pathway files for *E. coli*, and both the metabolic and non-metabolic pathway files for *S. cerevisiae* and breast cancer data sets. Each xml file was parsed using custom scripts, and the genes involved in the pathway were identified and used as the pathway genes. Those

pathways with fewer than five genes were removed from the analysis to avoid statistical insignificance. A total of 64, 88, and 188 *E. coli*, *S. cerevisiae*, and *H. sapiens* pathways met this criterion, respectively, and were used in this analysis.

3.1.4 Expression profiles

The microarray gene expression data were normalized using the Robust Multi-Array average expression measure (RMA) function from the *affy* package in R [40]. The expression profile, E_i, for gene i is represented as:

$$E_i = \{e_{i,1},\ldots,e_{im}\},$$

where e_{im} is the mean expression value of all probe sets for gene *i* on chip *m*. Gene expression profiles were created for each gene in a pathway, and each expression value was the $\log_2$ value for the normalized array intensity values.

In the breast cancer data set, noisy probes were removed. To accomplish this, those probes that were above the median of the chip in at least a quarter of the arrays were retained for the analysis. RMA was subsequently used to normalize the remaining probes.

3.1.5 Pathway correlation profiles

Pathway correlation profiles (e.g., correlation matrix) were created for each pathway in the data set. The profiles are calculated for all gene pairs among different chips at a given condition using Pearson correlations and are represented as:

$$\left\{\rho_{1,2},\rho_{1,3},\ldots,\rho_{i,j},\ldots,\rho_{k-1,k}\right\},$$

$$\rho_{i,j} = \frac{\sum_m (e_{im} - \mu_i)(e_{jm} - \mu_j)}{(n-1)\sigma_i \sigma_j},$$

where k is the number of genes in the pathway, m is the chip index, n is the total number of chips in a sample, μ_i and μ_j are the mean expression values for gene i and j, respectively, and σ_i and σ_j are the standard deviations, respectively. The pathway correlation profiles were calculated individually for each pH in *E. coli*, each time point in *S. cerevisiae*, and the ER-positive and ER-negative samples, respectively. Due to sample size biases in the breast cancer data set, the gene-gene pair correlations in the ER-positive class were estimated by repeatedly sampling 77 chips randomly and taking the final average of gene-gene pair correlations. Lastly, to ensure that the gene-gene pair correlations have a normal distribution and stable variance, the pathway correlation profiles were transformed using the Fisher transformation.

3.1.6 Pathway ranking

The derived pathway correlation profiles were used to rank the pathways based on most significant perturbation for each condition. In order to quantify the differences in correlation of specific gene-gene pairs between two conditions, the pathway perturbation was considered the average of these changes in correlation. Initially, a paired t-test was performed where each gene-gene pair correlation at one condition was directly compared to the corresponding correlation under the other condition. The paired t-test between condition (1) and condition (2) follows:

$$d = \left\{ F(\rho_{1,2}^{(1)}) - F(\rho_{1,2}^{(2)}), \ldots, F(\rho_{i,j}^{(1)}) - F(\rho_{i,j}^{(2)}), \ldots, F(\rho_{k-1,k}^{(1)}) - F(\rho_{k-1,k}^{(2)}) \right\}$$

where

$$F(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

and

$$t = \frac{\bar{d}}{s_{\bar{d}} / \sqrt{n}}$$

where d is the paired differences in Fisher transformed gene-gene pair correlations, $\bar{d}$ is the sample mean of the differences in transformed gene-gene pair correlations, $s_{\bar{d}}$ is the standard deviation of d, and n is the number of gene-gene pairs in the pathway. Due to a bias towards pathways of larger size, bootstrapping was then implemented separately in order to estimate the average change in gene-gene pair correlation for each pathway. For this, 100 gene-gene pairs were randomly sampled from each pathway and the average change in gene-gene pair correlations was calculated. From these samplings, the mean change in gene-gene pair correlations can be estimated using a *z-score.* We then combined the p-values from the Student's t-test analysis and bootstrapping, correct for multiple testing by using a Benjamini correction, and rank the pathways by corrected p-values. In particular, the p-values from the bootstrapping process remove the biases due to sizes of pathways. This method differs from the standard gene expression analysis in two ways: (1) we utilize all biological replicates as opposed to assessing the mean expression of individual genes, and (2) we calculate the changes in gene-gene correlation between conditions rather than calculating a change in expression between conditions.

3.2 Result: Rank pathways by *significance* of perturbation

3.2.1 Pathway ranking

Pathways were ranked based on their Benjamini corrected p-values that test if there is a significant change in the gene-gene pair correlations between two samples. Those pathways with a positive mean difference in correlations show that the gene-gene pair correlations from the treatment samples are on average higher than those gene-gene pair correlations derived from the normal samples. Conversely, those pathways with a negative mean difference show that there is a decrease in the pathway's correlation profile under the treatment conditions. **Table 1** and **Table 2** show the top ranking pathways from the *E. coli* data set when comparing pH 8.7 to an ideal pH 7, and the breast cancer data set when comparing ER-positive to ER-negative, respectively (full tables in Supplemental Tables 1-3).

There are 24 significantly perturbed pathways in the *E. coli* data set when looking at both pH 8.7 and pH 5 compared to the ideal pH 7 (Benjamini corrected p-value <0.01). The *S. cerevisiae* pathway mean difference and adjusted p-values were also calculated for the five time points compared to the control group (full results in Supplemental Table 2). There are 16, 23, 21, 23, and 19 significantly perturbed pathways when comparing the desiccation, at 0 (dry), 15, 45, 90 and 360 minutes, to the control group, respectively (Benjamini corrected p-value <0.01). In the breast cancer data set, the pathway correlation profile method identified 33 pathways as statistically perturbed when comparing ER-positive to ER-negative patient samples (Benjamini corrected p-value <0.01).

As a comparison, a Gene Set Enrichment (GSE) analysis using DAVID was also performed [41] and the most significant KEGG pathways [39] are reported in Table 1 and Table 2 for the *E. coli* and breast cancer data sets. Of the top 15 pathways reported from DAVID for the *E. coli* data, 20% overlapped with those deemed significantly perturbed from the pathway correlation profile analysis. This resulted in the pathway correlation profile method making numerous novel predictions for perturbed pathways that were not previously discovered using GSE methods such as DAVID. For the breast cancer data set, no pathways were found in common between the two methods.

Table 1. Comparison between DAVID Gene Set Enrichment Analysis and Pathway Correlation Profile analysis of E. coli pH data set at pH 8.7 compared to ideal pH 7.

Pathway Correlation Profile				DAVID Gene Set Enrichment		
KEGG ID	**Pathway Name**	**Mean Difference**	***p-value****	**KEGG ID**	**Pathway Name**	***p-value****
ecj00780	Biotin metabolism	0.549	6.46E-20	ecj00230	Purine metabolism	1.27E-07
ecj00523	Polyketide sugar unit biosynthesis	0.688	3.09E-16	ecj00190	Oxidative phosphorylation	2.05E-07
ecj01040	Biosynthesis of unsaturated fatty acids	0.590	2.02E-14	ecj00240	Pyrimidine metabolism	1.03E-06
ecj00230	Purine metabolism	0.186	1.71E-13	ecj00340	Histidine metabolism	3.53E-05
ecj00790	Folate biosynthesis	0.445	6.28E-07	ecj00020	Citrate cycle (TCA cycle)	1.96E-04
ecj00632	Benzoate degradation via CoA ligation	-0.516	6.89E-06	ecj00620	Pyruvate metabolism	1.99E-04
ecj00053	Ascorbate and aldar	0.452	5.23E-05	ecj00500	Starch and sucrose metabolism	4.36E-04
ecj00380	Tryptophan metabolism	-0.547	7.69E-05	ecj00670	One carbon pool by folate	2.81E-03
ecj00900	Terpenoid backbone biosynthesis	0.462	4.54E-04	ecj00650	Butanoate metabolism	4.10E-03
ecj00471	D-Glutamine and D-glutamate metabolism	0.413	5.35E-04	ecj00040	Pentose and glucuronate interconversions	4.20E-03
ecj00020	Citrate cycle (TCA cycle)	-0.246	5.85E-04	ecj00010	Glycolysis / Gluconeogenesis	6.74E-03
ecj01053	Biosynthesis of siderophore group nonribosomal peptides	0.460	5.85E-04	ecj00030	Pentose phosphate pathway	7.65E-03
ecj01110	Biosynthesis of secondary metabolites	0.039	6.64E-04	ecj00052	Galactose metabolism	1.10E-02
ecj00740	Riboflavin metabolism	0.526	1.05E-03	ecj00632	Benzoate degradation via CoA ligation	2.27E-02
ecj00450	Selenoamino acid metabolism	0.433	1.12E-03	ecj00250	Alanine, aspartate and glutamate metabolism	3.03E-02

Pathway rankings based on adjusted p-values. Those pathways with positive mean differences show that the gene-gene pairs on average have a higher correlation at a stressed pH, and a lower correlation at an ideal pH. *: Benjamini correction.

Table 2. Comparison between DAVID Gene Set Enrichment analysis and Pathway Correlation Profile analysis of the human breast cancer data set.

Pathway Correlation Profile				DAVID Gene Set Enrichment		
KEGG ID	Pathway Name	Mean Difference	p-value*	KEGG ID	Pathway Name	p-value*
hsa03040	Spliceosome	-0.0412	5.83E-30	hsa05219	Bladder cancer	0.004
hsa04080	Neuroactive ligand-receptor interaction	-0.0367	6.48E-22	hsa05200	Pathways in cancer	0.024
hsa03010	Ribosome	0.0382	2.74E-17	hsa04110	Cell cycle	0.069
hsa04514	Cell adhesion molecules (CAMs)	0.0298	1.63E-12	hsa04062	Chemokine signaling pathway	0.076
hsa00061	Fatty acid biosynthesis	-0.1712	3.32E-12	hsa04115	p53 signaling pathway	0.075
hsa00982	Drug metabolism - cytochrome P450	-0.058	4.79E-09	hsa00380	Tryptophan metabolism	0.063
hsa00140	Steroid hormone biosynthesis	-0.088	4.49E-08	hsa05215	Prostate cancer	0.058
hsa04060	Cytokine-cytokine receptor interaction	0.0152	5.41E-07	hsa05222	Small cell lung cancer	0.087
hsa05330	Allograft rejection	-0.0452	7.69E-07	hsa00010	Glycolysis / Gluconeogenesis	0.400
hsa00980	Metabolism of xenobiotics by cytochrome P450	-0.059	8.55E-07	hsa04144	Endocytosis	0.387
hsa03050	Proteasome	-0.0451	9.40E-07	hsa04512	ECM-receptor interaction	0.365
hsa00232	Caffeine metabolism	-0.1427	1.32E-06	hsa04114	Oocyte meiosis	0.375
hsa04740	Olfactory transduction	0.0395	1.50E-06	hsa04510	Focal adhesion	0.362
hsa05322	Systemic lupus erythematosus	0.0171	3.46E-06	hsa04960	Aldosterone-regulated sodium reabsorption	0.349
hsa04142	Lysosome	-0.0174	2.86E-05	hsa00330	Arginine and proline metabolism	0.360

Pathway rankings based on adjusted p-values. Those pathways with positive mean differences show that the gene-gene pairs on average have a higher correlation in ER-positive patient samples and a lower correlation in ER-negative patient samples for that pathway. *: Benjamini correction.

3.2.2 Pathways perturbed in E. coli

After our pathway correlation profile analysis, the *E. coli* metabolic pathways were ranked by p-values (the top 15 significant pathways shown in **Table 1**; full pathway results provided in Supplemental Table 1). Those pathways with a mean difference greater than zero show an increase in gene-gene pair correlations under the stressed pH 8.7 when compared to an ideal pH 7.

When comparing *E. coli* at pH 8.7 against the ideal pH 7, a majority of the significant pathways (19 out of 24) show an increase in gene-gene correlations during the basic environment (p-value < 0.01). Similarly, when comparing pH 5 to the ideal pH 7, only 18 pathways out of the 24 significantly perturbed pathways show this increase in gene-gene correlations under these conditions. In fact, only 14 pathways in common are significant under both stressed conditions, when compared to the ideal pH.

The Biotin Metabolism pathway (ecj00780) was the top ranked pathway based on perturbation when comparing the samples at pH 8.7 with those at pH 7. The kernel smoothed density graphs of the pathway correlation profiles at pH 5, pH 7, and pH 8.7 from the Biotin Metabolism pathway are shown in Figure 3a. The pathway correlation profile at pH 8.7 (in blue) shows an overall increase in untransformed gene-gene pair correlations within the pathway, suggesting a convergence towards a more consistent profile of the pathway during this stress. There is minimal difference between the pathway correlation profiles for this pathway at pH 5 and pH 7, also supported by the non-significant p-value. In the analysis, the Fisher transformed pathway correlation

profiles for the Biotin Metabolism pathway under each condition were directly compared (Figure 3b).

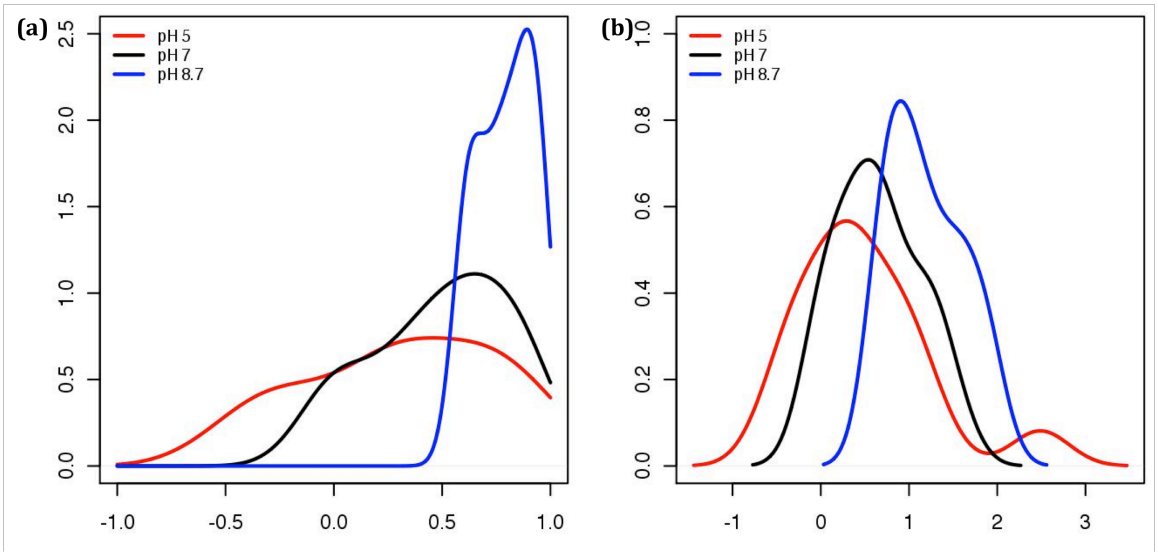


Figure 3. Pathway correlation profiles for Biotin Metabolism Pathway (ecj00780) in E. coli.

(a) Pathway correlation profile kernel density smoothed graphs before fisher transformation of the Biotin Metabolism Pathway. **(b)** Pathway correlation profile kernel density smoothed graphs after fisher transformation of the Biotin Metabolism Pathway. (pH 8.7: blue, pH 7: black, and pH 5: red).

3.2.3 Pathways perturbed in *S. cerevisiae*

Our pathway correlation profile method compared the *S. cerevisiae* treatments (desiccated and four rehydration time points) to the control samples. Both the metabolic and non-metabolic pathways were ranked based on Benjamini corrected p-values. Complete final pathway results are in Supplemental Table 2. When looking at the time series data, a few trends can be identified. Three pathways show a statistically significant

increase in correlation while desiccated, and no significant changes in gene-gene correlation throughout rehydration. Six pathways show the most significant decreases in gene-gene correlation after 360 minutes of rehydration, and less significant decreases in pathway correlation at all other time points. These pathways, including the DNA replication and mRNA Surveillance pathway, show a trend towards an uncorrelated state as the rehydration process progresses (i.e. more negative mean difference). Seven pathways show significant perturbation at all time points when compared to the control sample. Of these, only the Ribosome pathway shows a convergence towards a more correlated state during all the time points. The untransformed pathway correlation profiles for the Ribosome pathway (sce03010) are shown in **Figure 4**b. The profile for the control sample shows a more random distribution of correlations; whereas the profiles for all the time points of rehydration and the desiccated sample show a strong skew towards a highly correlated state. **Figure 4**c demonstrates the pathway correlation profile distributions approaching normal after the Fisher transformation.

3.2.4 Pathways perturbed in Breast Cancer

Our pathway correlation profile analysis was performed on the ER-positive/ER-negative breast cancer data set and both the metabolic and non-metabolic pathways were ranked by adjusted p-values (top 15 significant pathways shown in **Table 2**; full pathway results provided in Supplemental Table 3). In total, 33 out of 188 pathways were ranked as significantly perturbed (Benjamini corrected p-value < 0.01) and 70% of these pathways show increases in gene-gene correlation in the ER-negative patient samples when compared to those from ER-positive patient samples.

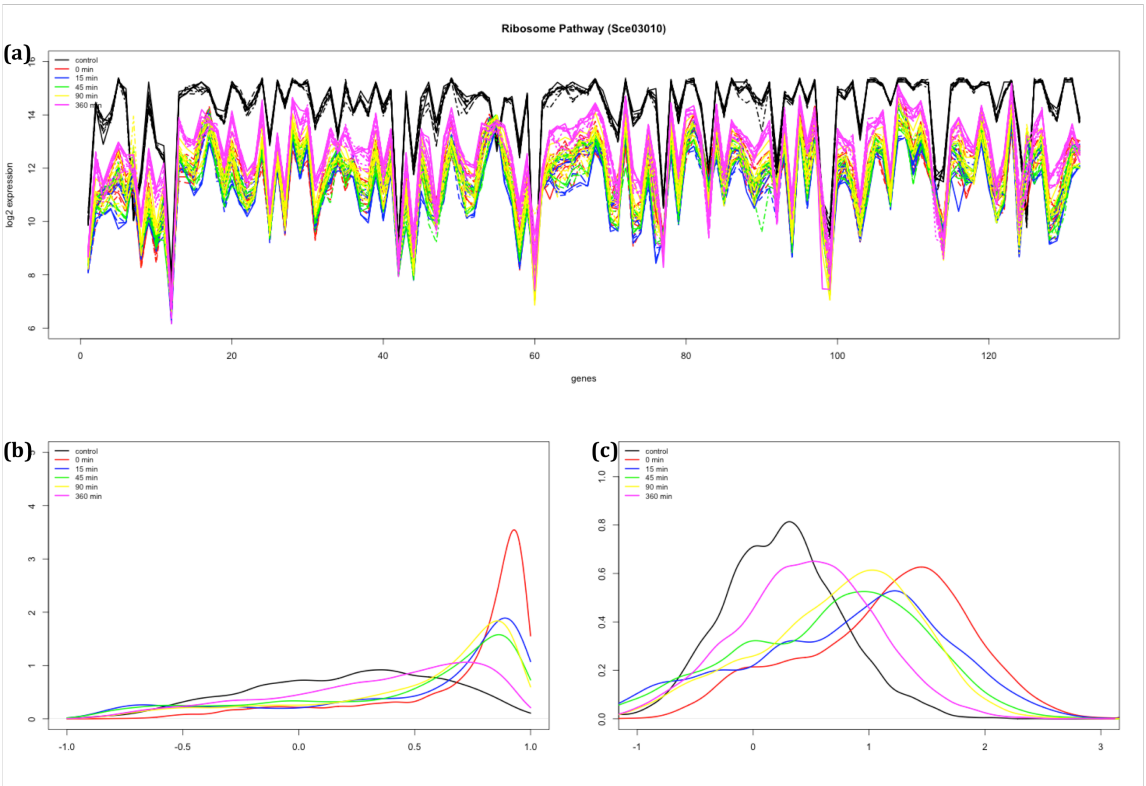


Figure 4. Pathway correlation profiles for Ribosome Pathway (sce03010) in *S. cerevisiae*.

(a) Gene expression level plots of the Ribosome Pathway **(b)** Pathway correlation profile kernel density smoothed graphs before fisher transformation of the Ribosome Pathway. **(c)** Pathway correlation profile kernel density smoothed graphs after fisher transformation of the Ribosome Pathway. (Control: black, 0 minutes: red, 15 minutes: blue, 45 minutes: green, 90 minutes: yellow, and 360 minutes: magenta).

The Spliceosome pathway (hsa03040) and the Neuroactive Ligand-Receptor Interaction pathway (hsa04080) were ranked as most perturbed when comparing the receptor status groups. Both of these pathways show an average decrease in gene-gene pair correlations when comparing ER-positive to ER-negative patient samples. Due to the larger variations among patients in cancer data sets than those in single-cell microbes (*E. coli* and *S.*

cerevisiae), the gene-gene pair correlations are smaller in magnitude. Thus, the mean difference of the pathway is smaller, when compared to more controlled data sets of *E. coli* and *S. cerevisiae*.

3.3 Discussion

Our pathway correlation profile method relies on the assumption that non-significant genes/pathways have a random correlation as their background "noise." To show that a pathway is perturbed (i.e. activated or repressed), we need to show that the pathway no longer maintains a random gene-gene correlation profile but rather takes on a more convergent profile. This convergence could be either towards a coherently regulated state, as indicated by positive changes in correlations, or a dysregulated state, as indicated by negative changes in correlations. Final interpretations of the correlation profiles are likely to depend on the gene expression trends within the pathway.

A majority of the top 15 most significant *E. coli* pathways under basic conditions, when ranked by p-value, show an increase in correlation under the extreme conditions. This increase in correlation suggests that those pathways have a more consistently regulated system under these conditions when compared to normal; hence these pathways are likely to be universally activated or repressed in a highly coordinated manner. As for the acidic conditions, a majority of the 24 significant pathways show an increase in pathway correlation profiles, suggesting many pathways require activation/coordination of their expression under stressful conditions.

The increase in gene-gene pair correlations of the Biotin Metabolism pathway in *E. coli* at a pH of 8.7 (Figure 3), when compared to the ideal pH 7, suggests this pathway is activated at pH 8.7 and converges to a more correlated state. Biotin is a relatively unstable molecule in alkaline conditions [42], and in *E. coli* the majority of the genes in the Biotin Metabolism pathway are part of the bio-operon [43]. With a decrease in the stability, and presumably therefore the abundance of biotin under alkaline conditions, there is increased expression of the bio-operon. Since these genes are organized as an operon, increase in expression of one gene results in a coordinated increase in expression of the other genes, which can be quantified through increases in gene-gene correlations. This is confirmed by the results from the pathway correlation profile method which show an increase in gene-gene pair correlations at pH 8.7 compared to pH 7 (Figure 3a,b, **Figure 6**). Moreover, there is an increase in gene expression at pH 8.7, when compared to the other pHs (**Figure 5**).

In the *S. cerevisiae* data set, there were varying changes in pathway correlation profiles throughout the time points within the experiment. At the time of desiccation, 16 pathways had significantly different pathway correlation profiles from those in the control (corrected p-value < 0.01). Of these, eight pathways have a decrease in gene-gene correlations, and eight show an increase in gene-gene correlations, including the Ribosome pathway (sce03010) and Cell Cycle in Yeast (sce04111). Of these eight pathways, only two show this significant increase in pathway correlation profiles at the time of desiccation and no significant changes during rehydration (Cell Cycle in Yeast and Nucleotide Excision Repair). These two pathways, in essence, show increases in consistency of regulation at the time of desiccation with subsequent non-regulation

during rehydration. This suggests that these pathways were necessary for cell survival at the time of desiccation but were not necessary throughout the rehydration process.

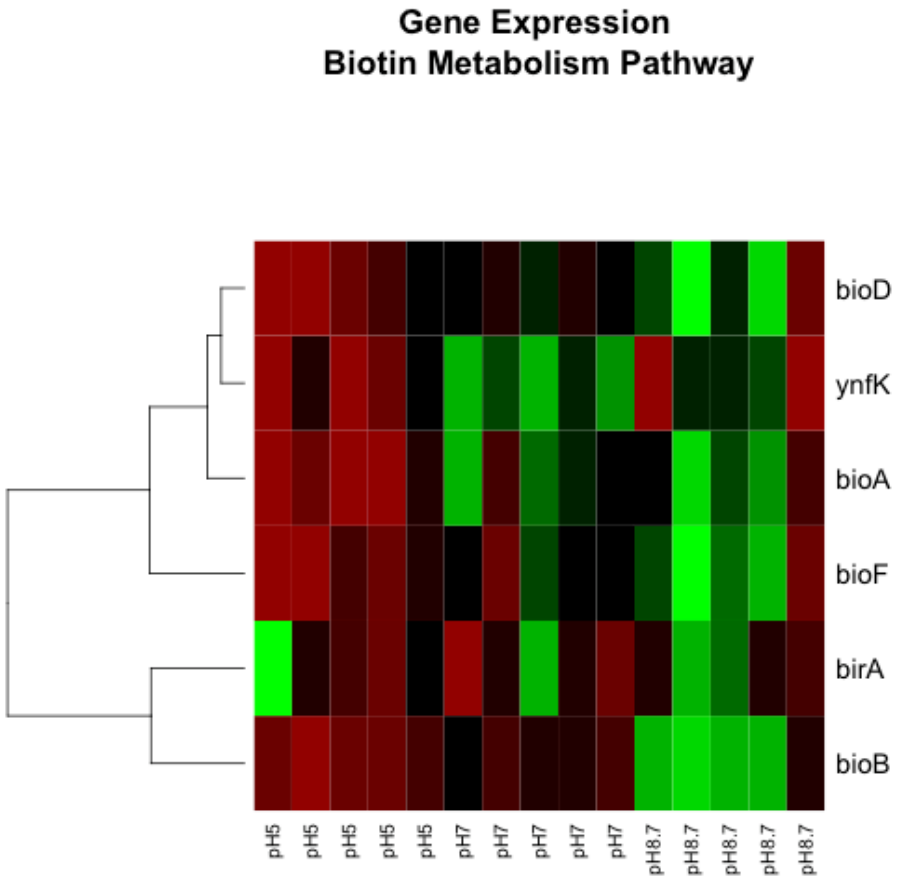


Figure 5. Heatmap of the gene expression for those genes in the Biotin Pathway (ecj00780) in E. coli under three pH conditions.

Heatmap and clustering of genes are based on gene expression signal. Rows are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; red: negative z-normalized expression level)

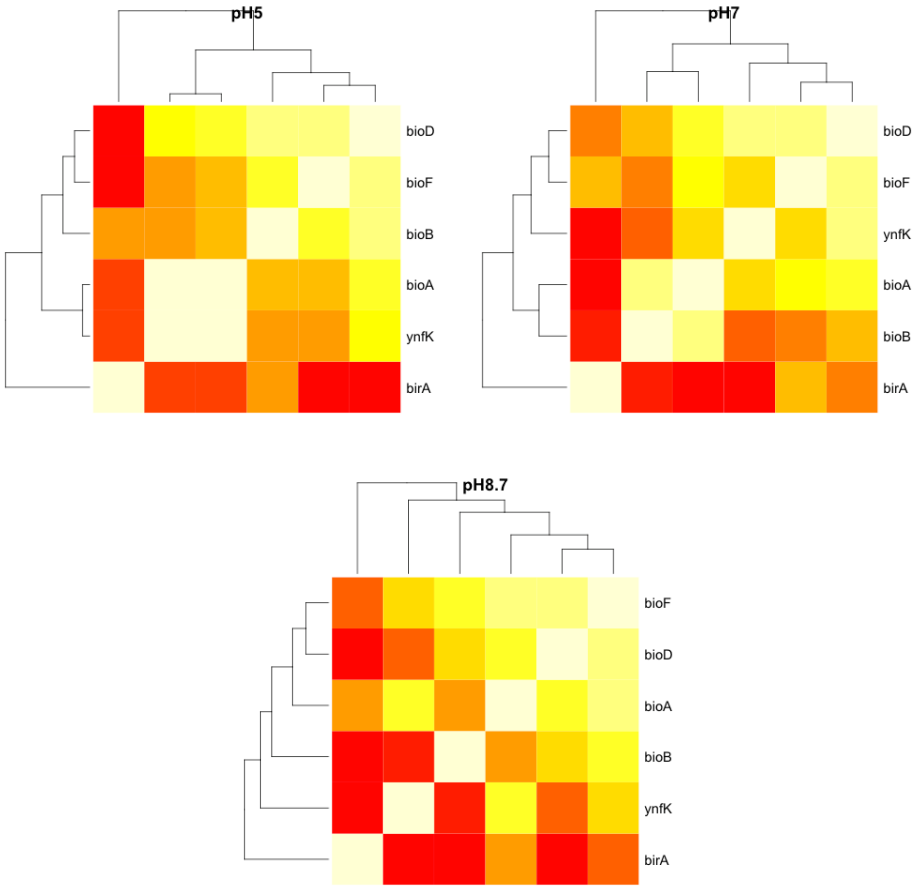


Figure 6. Heatmap of pathway correlation profiles for Biotin Metabolism Pathway (ecj00780) in E. coli under each of the three pH conditions, respectively.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows and columns represent genes. (Yellow: higher correlation values; red: lower correlation values. Note: the red-yellow range is relative to each individual heatmap. See Figure 3 for reference on ranges of correlation values.)

A significant increase in pathway correlation suggests:

1. the cells in the sample taken are a more homogeneous set of cells than those in the control sample; and/or
2. the pathway shows a more stable and consistent expression profile among the genes involved in this pathway, such as a regulated pathway.

These hypotheses can be shown through the Yeast Cell Cycle and the Ribosome pathways, respectively.

A more homogenous population of cells will reduce the biological variation in gene expression from positive signals, i.e., there is a stronger relationship between gene expression and phenotype [44]. According to Singh et al. the cells remained in the G_0/G_1 phase, or in a "holding pattern" at the time of desiccation and throughout the rehydration process [45]. The increase shown in the Cell Cycle pathway correlation at the time of desiccation suggests that there is an increase in pathway regulation stemming from a decrease in gene expression variation within this pathway. These results, together with the coincidence in cell cycle timing, suggest a more homogeneous population of cells.

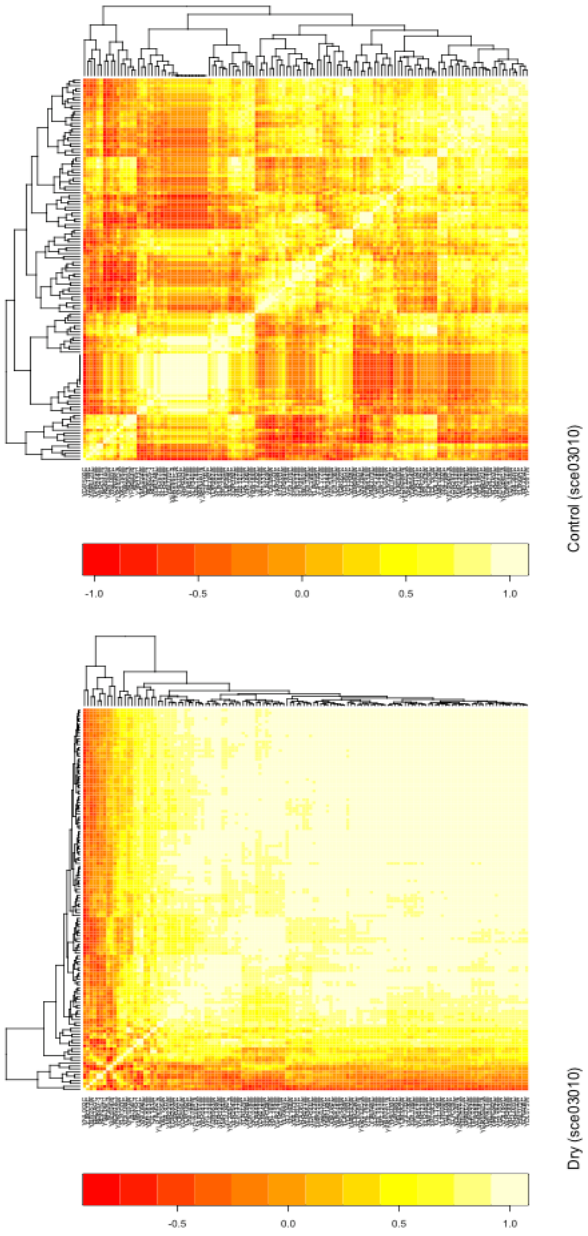
Genes that are working together show increases in gene co-expression and coalesce into a more synchronous pathway [46]. The Cell Cycle pathway shows this more stable and consistent pathway correlation profile among the genes involved in this pathway. Besides showing a significant increase in correlation while desiccated, this pathway shows no significant changes at the onset of rehydration; however, it then shows a progressive move towards convergence to a more correlated state as the rehydration processes

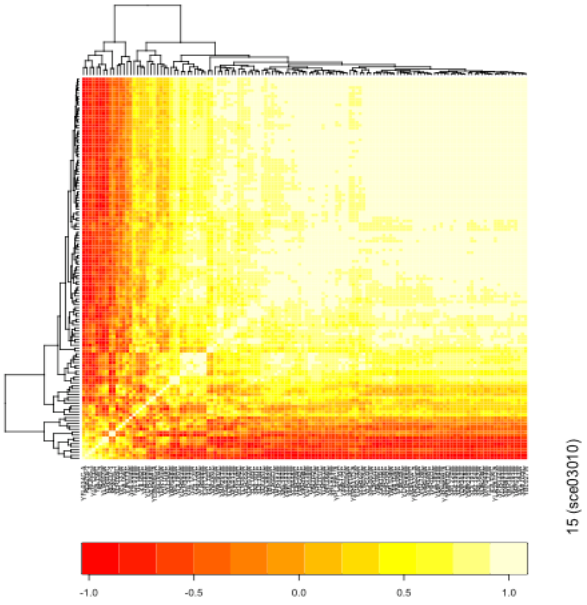
progressed, though still not significant. During this "holding pattern" time from desiccation through rehydration, the Cell Cycle pathway may not need to be orchestrated since the cells do not progress through the cell cycle. Instead, these genes show a more random background profile as would be expected from an unregulated pathway.

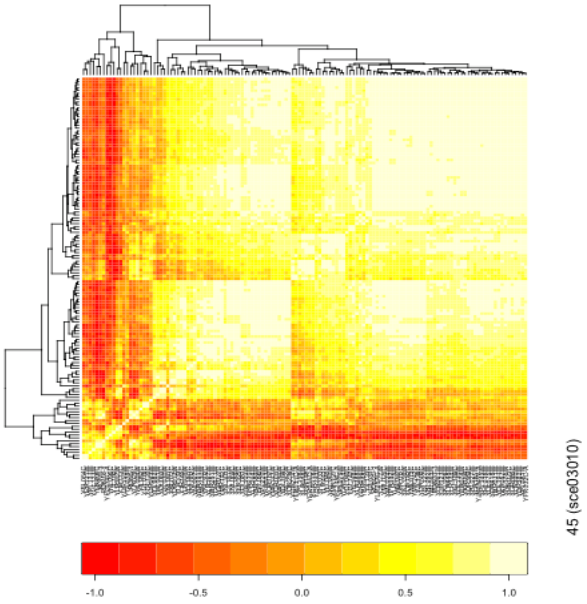
In contrast to the increase in pathway correlation of the Cell Cycle pathway, the DNA Replication pathway (sce03030) shows a significant decrease in gene-gene correlations at desiccation and at subsequent rehydration time points. Given that the cell population at these time points is held in the G_0/G_1 phase and not the S phase, there is no DNA replication occurring. All of this taken together suggests that the DNA Replication pathway is not essential for cell survival during the desiccation and rehydration, and is therefore dysregulated.

The Ribosome pathway (sce03010), on the other hand, shows an increase in pathway correlation profile at all time points compared to the control. These positive correlation changes suggest that the pathway is regulated at all time points. This regulation (deactivation) can be shown through the strong, and nearly universal, decreases in gene expression within this pathway (**Figure 4**a). For all time points, including desiccation and throughout rehydration, the 132 genes in this pathway show a decrease in expression, averaging greater than 4-fold change when compared to the control sample's expression profile. Given this pathway is regulated at all time points, identifying modules/clusters of genes that are coordinately regulated at each time point is important in understanding the pathway dynamics. Using a heatmap of the gene-gene pair correlations (**Figure 7**), we can cluster the genes at each time point. Through these heatmaps, we can show that there

are strong and dynamic clusters of genes that co-express together at each particular time point suggesting varying modules of regulation that are differentially activated at each time point.




15 (sce03010)
-1.0
-0.5
0.0
0.5
1.0


45 (sce03010)
-1.0
-0.5
0.0
0.5
1.0

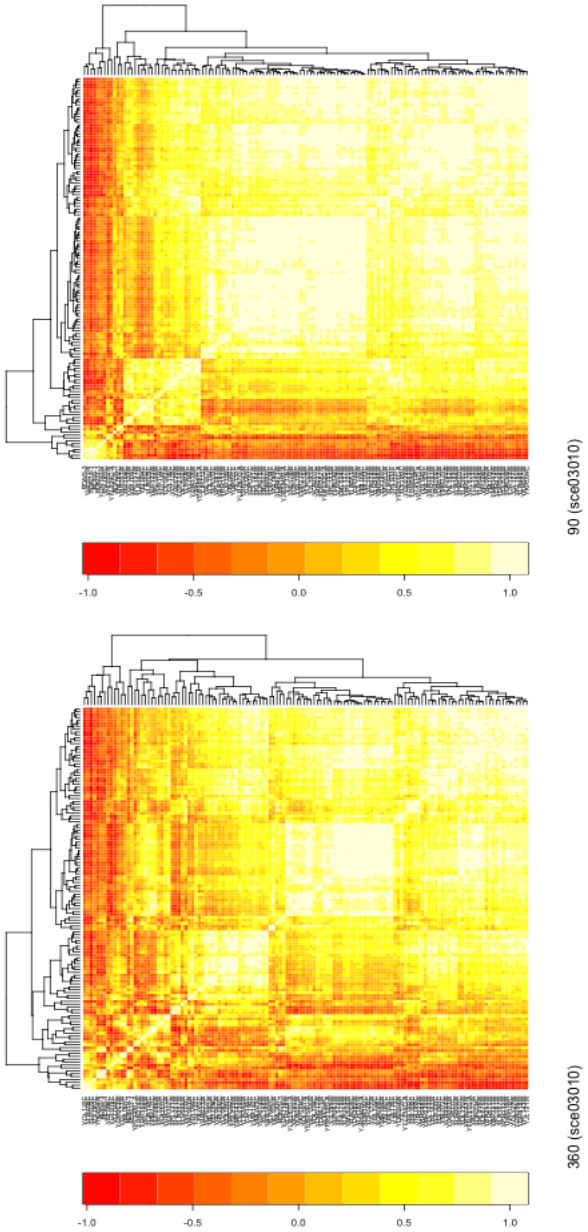


Figure 7. Heatmap of pathway correlation profiles for the Ribosome Pathway (sce03010) in *S. cerevisiae* under each of the three pH conditions, respectively.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows and columns represent genes. (Yellow: positive correlation; red: negative correlation)

Just over half of the *S. cerevisiae* pathways, however, showed no significant changes in pathway correlation profiles throughout the desiccation and rehydration process (based on corrected p-values). These pathways, including many metabolic pathways, are likely to have not changed in functional regulation or are not essential for the cell's survival during these times. It follows that the metabolic pathways are necessary to function while the cell is still alive, and drastic changes in these pathways could result in cellular death.

In contrast, 7 out of 88 pathways (8%) show a significant change in gene-gene correlation during all time points of the experiment when compared to the control sample. All of these pathways, except the Ribosome pathway, show a decrease in gene-gene correlation throughout the desiccation and rehydration process. This decrease in gene-gene pair correlations could infer that these pathways are necessary for cell survival and/or proliferation under normal conditions, but once stresses are induced, these pathways are no longer required to be regulated during duress.

In comparing our method to the well accepted DAVID gene set enrichment method [41], there was some concordance between results on the *E. coli* pH 8.7 data. With 20% of the top 15 pathways in common, our method identifies not only significant pathways that would have been previously discovered given these experimental conditions, but also uncovers 12 additional pathways that would not previously have been investigated. The gene-gene pair correlations allow for an alternative perspective on pathway perturbation and the utilization of biological replicates independently, therefore identify significant pathways through different assumptions. The folate biosynthesis pathway (ecj00790), one of these 12 novel predictions, shows an increase in gene-gene pair correlations when

comparing pH 8.7 to the control pH for *E. coli* (**Figure 8**,**Figure 9**). Within this pathway, a majority of the genes show small increases in gene expression under the basic conditions. Due to none of these increases in gene expression being statistically significant, this pathway was not reported in DAVID, a standard GSE method. It has been shown in selected species of lactic acid bacteria that higher pHs allow for increases in folate levels, suggesting more efficient folate biosynthesis under these conditions [47]. Similar explanations about folate biosynthesis could be inferred for *E. coli* under basic conditions. By exploring the data through a different perspective, i.e., our pathway correlation profiles, we can identify new pathways that have the potential to be involved in the condition and further add insight into explaining the biological mechanisms that occur within the cell when stressed at pH 8.7.

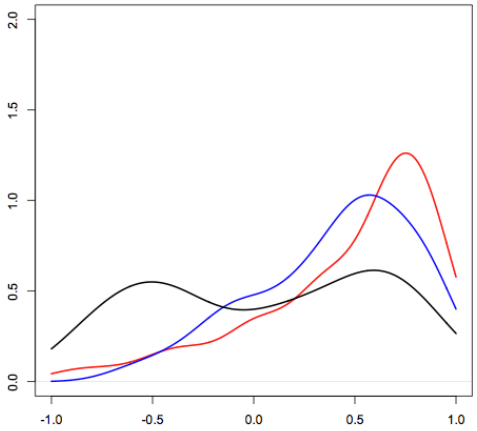


Figure 8. Pathway correlation profiles for the Folate Biosynthesis Pathway (ecj00790) in E. coli.

Pathway correlation profile kernel density smoothed graphs before fisher transformation of the Folate Biosynthesis Pathway. (pH 8.7: blue, pH 7: black, and pH 5: red).

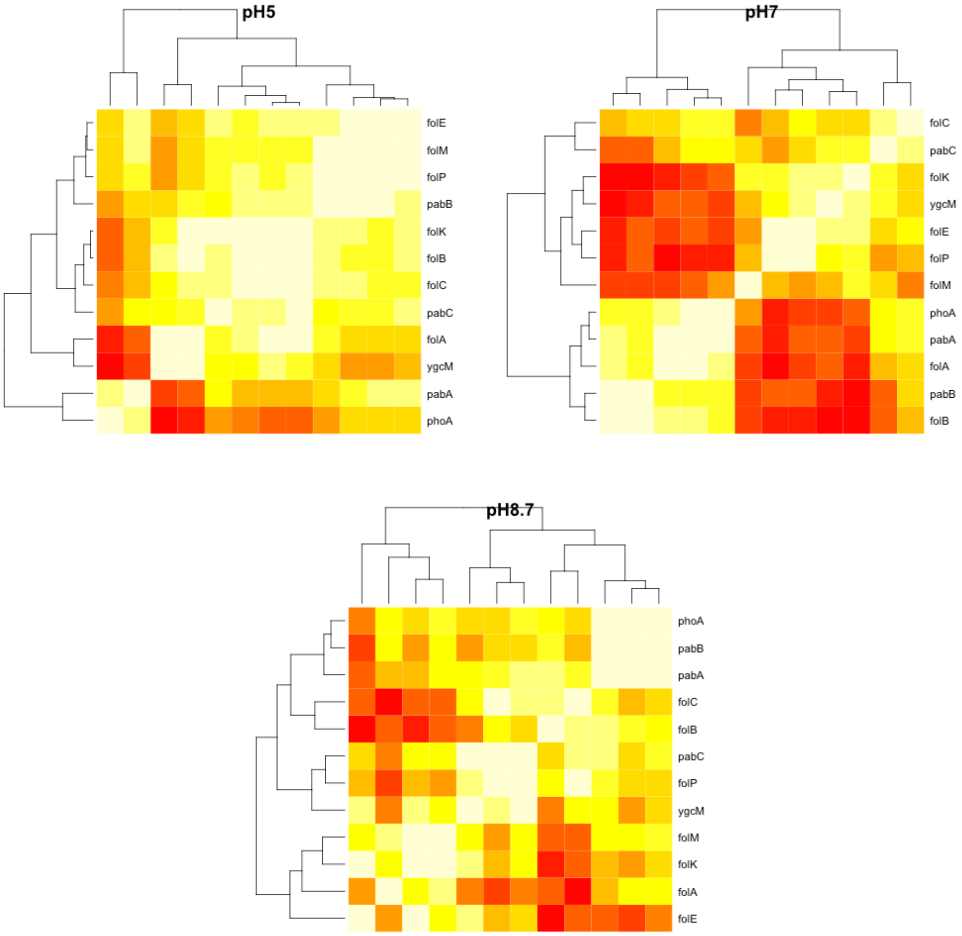


Figure 9. Heatmap of pathway correlation profiles for the Folate Biosynthesis Pathway (ecj00790) in E. coli under each of the three pH conditions, respectively.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows and columns represent genes. (Yellow: higher correlation values; red: lower correlation values. Note: the red-yellow range is relative to each individual heatmap. See **Figure 8** for reference on ranges of correlation values.)

The pathway correlation profile method was used to analyze a breast cancer estrogen receptor data set. When comparing the results from our method with those from the DAVID gene set enrichment method, no pathways were found in common. In fact, only

two pathways were deemed significant when using DAVID (Benjamini adjusted p-value < 0.01). The discordance in predictions between our method and the DAVID method is due to the different assumptions regarding pathway perturbation as well as the DAVID method having a bias towards larger pathways, whereas our method tries to reduce pathway size biases. As a result, we can make predictions of perturbation that are independent of size.

Wang et al. reported that their gene signature for differentiating ER-positive from ER-negative patients included pathways involved in cell death, cell cycle and proliferation, DNA replication and repair, and immune response [38]. The pathway correlation profile did in fact find perturbation in the DNA Replication pathway and the Cell Cycle pathway (p-value <0.05; Supplemental Table 3), and so did DAVID. The pathway correlation profile method found the Neuroactive Ligand-Receptor Interaction pathway (hsa04080) perturbed in breast cancer, but DAVID found this pathway non-significant. Within this pathway, PTGER3 is involved in many of the largest changes in gene-gene correlations. Though PTGER3 has minimal change in gene expression, the average change in Fisher transformed gene-gene correlations between this gene and all other genes in the pathway is 0.34, with increases in gene-gene pair correlations in ER-negative patient samples. Further validation is needed to show the relation between PTGER3 and estrogen receptor status in breast cancer.

Here, we have used a pathway correlation profile method to identify perturbed pathways in *E. coli, S. cerevisiae,* and a human breast cancer data set. Our method takes a global approach to analyzing gene expression data for identifying pathway perturbation. First,

we take advantage of the prior knowledge of pathway members and use this to efficiently and effectively analyze the data. Second, we no longer rely on single gene involvement to identify significant pathways; rather, we look at the overall relationship between genes within a pathway and determine the level of perturbation based on changes in gene-gene relations, regardless of a specific gene's expression profile. Third, our method exploits the biological repeats of gene expression data, while existing methods often take an average of the repeats without using the data explicitly. Lastly, our method is more robust and less influenced by the inherent noise that comes from microarrays. This method can also be adapted for additional pathway databases, such as Reactome [48], TRANSPATH [49], and pre-defined gene ontologies, as well as alternate data platforms, such as RNA-seq [50].

4 Chapter 4: Ribosome Pathway

4.1 Introduction

The Ribosome is essential to cell survival; this protein complex is responsible for the translation process from RNA to protein [51]. For all species, this complex is made up of two subcomponents, a large and small subunit. In the ribosomal complex, there are three binding sites that facilitate protein translation: the aminoacyl, peptidyl, and exit [52]. Through these sites, the ribosome, mRNA and tRNA work together to accurately translate the protein as well as assist in the protein translocation. When comparing prokaryotes and eukaryotes, the structure of the ribosome complex have minimal structural differences [51]. Exploiting these small deviations between species-specific ribosomes provide weaknesses ideal for targeted therapeutics, for example antibiotics [53,54,55].

By gaining a more global understanding for the regulation of the Ribosome genes, we can learn new insights into possible mechanisms of cell survival. With these insights, we can approach disease understanding from a new perspective, identify potential targets for therapeutics, and decipher how some samples/species survive under selective conditions while others do not.

It is not a new concept to consider pathways evolving in concordance with species evolution, especially with respect to metabolic pathways [56,57,58,59,60]. Similar studies have also looked at signaling pathway adaption over time [61]. In most instances, these pathways evolve in a modular fashion. Maintaining groups of genes that evolve together allows for a more coordinated method of regulation. By expanding the idea of

pathway evolution via this modular structure, we can utilize these units of regulation as a means to explain pathway dynamics for different tissues and experimental conditions. Moreover, we can use our pathway correlation profiles to help identify these modules, or sub-pathway structures throughout various tissues and experimental conditions.

Here, we used the developed pathway correlation profiles in order to characterize the Ribosome pathway under various normal tissues followed by experimental conditions in *E. coli*. We hypothesize that there is tissue/condition specific regulation of the Ribosome Pathway, thereby resulting in a module structure of regulation to the pathway. It is not an unfamiliar concept to consider tissue specific regulation of the Ribosome complex, as previously identified by Ramagopal and Ennis [62]. This approach results in the identification of sub-pathways which can better explain the adaptations of the Ribosome complex in differential tissues and conditions.

4.2 Data

To better understand the Ribosome Pathway, we have used data from two species, human and *E. coli*. Six human data sets of normal tissue types were downloaded from the GEO website and used in a meta-analysis to characterize the human Ribosome Pathway: skin [63], breast [64], brain [65], b-cell [66], ovary [67] and kidney [68]. The skin data set had 13 samples and the experiment investigated differences between normal skin tissue and psoriatic plaques. The breast data set looked into changes in gene expression patterns of normal tissue samples taken from tumor patients and cancer-free prophylactic mastectomy patients, and the 18 normal samples from the cancer-free prophylactic mastectomy patients were used. For the brain tissue data set, 31 samples were collected

post mortem. Lastly, the b-cell, ovary and kidney had 11, 4, 5 samples, respectively, and were each from experiments that interrogated their corresponding cancers.

All these data sets used the Affymetrix HG-U133A platform, and were analyzed independently of each other. Each data set was individually processed using RMA to normalize the gene expression and resulted in log expression values. The expression data was then used to derive the pathway correlation profiles for each of the KEGG pathways.

4.3 Results and Discussion

Understanding a pathway starts with looking at the gene expression of those genes within the pathway. A heatmap of the gene expression for the Ribosome Pathway is shown in **Figure 10**. When completing the hierarchical clustering on tissue types, all the samples from one tissue type cluster together, with the exception of one kidney sample that clusters with brain. On average, the skin samples have higher expression levels of all Ribosomal genes than any other tissue type. This increase in skin ribosomal gene expression, when compared to the other tissue types, could be a result of the increased turnover rate of skin cells. Since skin is a constantly renewing organ, these cells will consistently need to transcribe and translate genes, hence the increase in expression of the Ribosomal genes throughout cell growth and differentiation.

In contrast, the brain samples have decreased gene expression of most all of the Ribosomal genes, when compared to the other tissues. In the experiment for the normal brain tissue, samples were taken post mortem. As a result, these samples are very likely to be inactive cells that have minimal to no signaling occurring within. If there is minimal cellular activity, there will also be decreases in gene expression and translation. Thereby,

an increase in the expression of the Ribosomal genes is not necessary, since no new Ribosome complexes need to be established.

However, the ovary, b-cell, and breast samples show varying up and down regulated genes within the Ribosome Pathway. Based on the sample hierarchical clustering, there is a tissue specific expression profile for those genes involved in the Ribosome pathway. For these tissue types, selective genes are up-regulated and down-regulated in each tissue, when compared across all tissue types. This is different from the skin tissue that was universally up-regulated and the brain tissue that was universally down regulated. Due to the varying direction of the gene expression, this possibly suggests that there is a tissue specific control over the expression of these genes. Alternatively, tissue specific genes could exist in the Ribosome pathway.

Having tissue specific gene expression is not novel, as shown by Kondrashov et. al [69], where they identified specificity of gene expression for various tissues due to a regulation of ribosomal proteins. The tissue specific expression patterns seen in the ovary, b-cell and breast tissue help to support the concept that ribosomal proteins are regulated, in order to impose additional post-transcriptional regulation on the system. When looking at the gene expression levels individually, we are able to determine that each tissue has its own expression pattern, but how and why these expression patterns change between tissues is still elusive. Moreover, gene expression alone is not responsible for pathway regulation, and in many instances a clear gene expression pattern for a pathway cannot be identified even for different tissues. An alternate perspective would be to assess the profiles of the gene-gene pair correlations for a given tissue type.

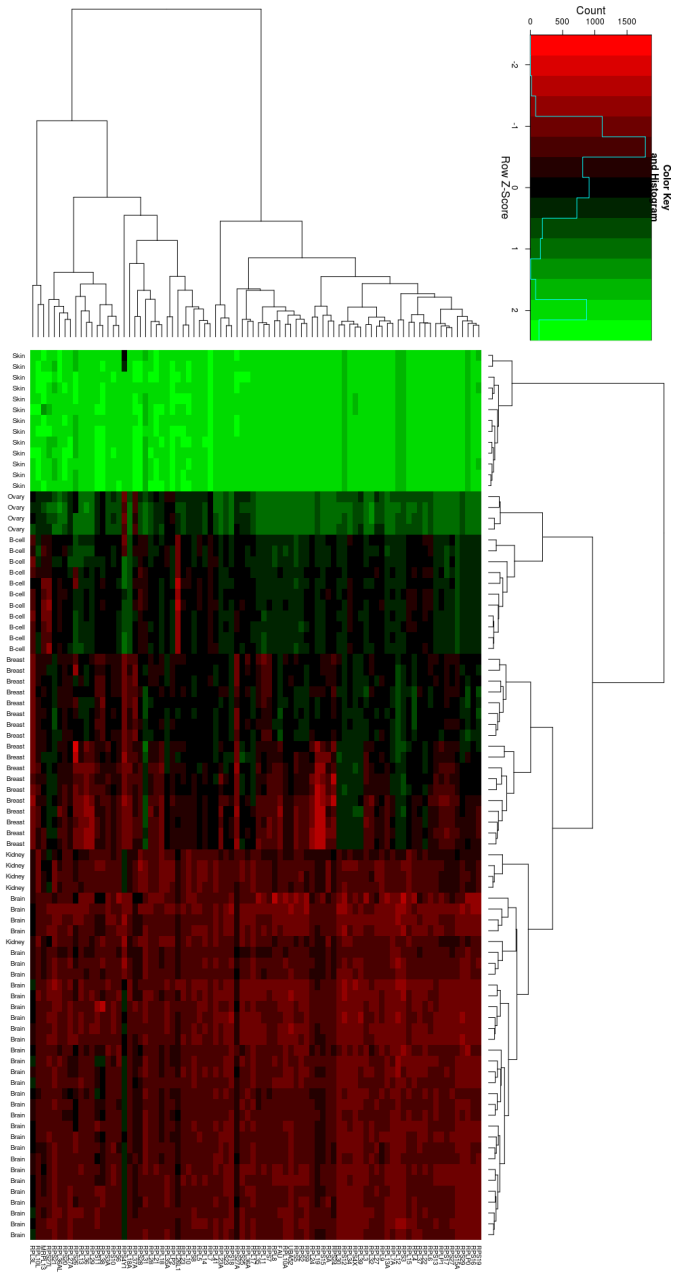
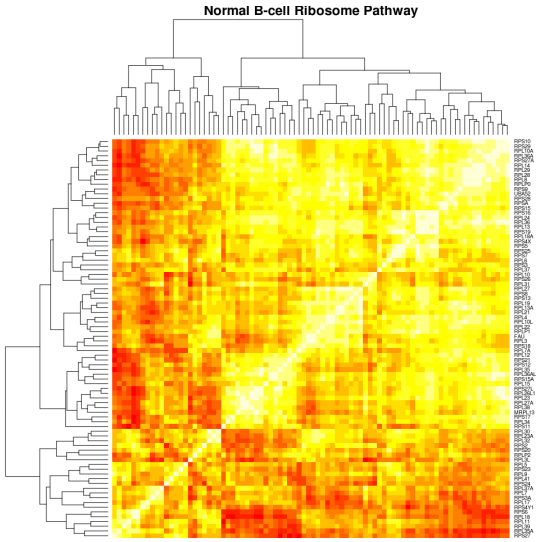


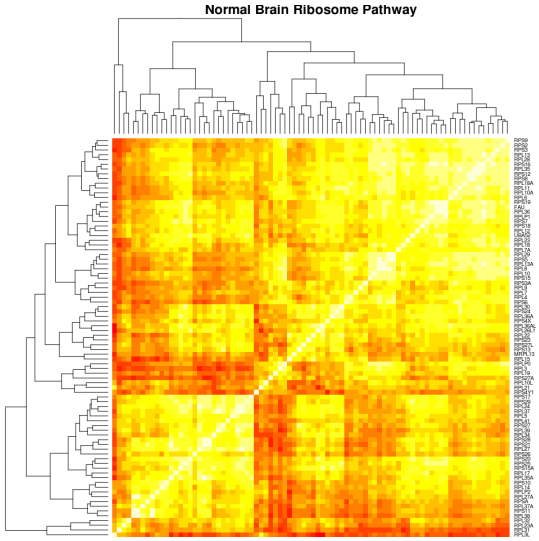
Figure 10. Heatmap of the gene expression for those genes in the Ribosome Pathway (hsa03010) in humans for various normal tissue data sets.

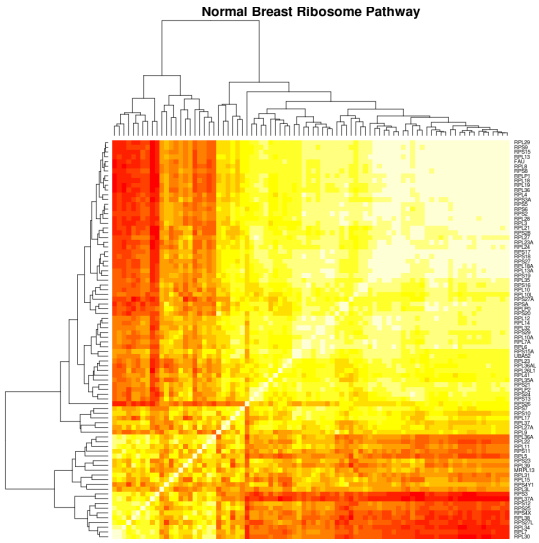
Heatmap and clustering of genes are based on gene expression signal. Rows are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; red: negative z-normalized expression level)

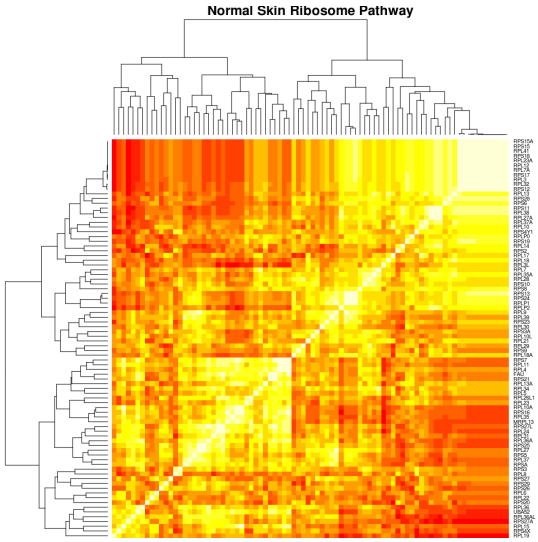
By looking at the heatmaps for the pathway correlation profile of each of the normal tissues (**Figure 11**), several conclusions can be inferred. In addition to each tissue having a distinct expression profile, each tissue has its own correlation profile. The kidney samples have mostly positive gene-gene pair correlations and a few select genes that are negatively correlated with all the other genes except themselves. A pattern like this could suggest (1) those few select genes that are negatively correlated with majority of the pathway are part of a repressive complex that acts on the entire pathway, or (2) those few select genes are not expressed in the kidney tissue, and as a result have a negative gene-gene pair correlation with the rest of the Ribosomal genes that are expressed in the kidney tissue. Either of the two options suggests that there is a specific type of Ribosomal regulation for the kidney tissue, which is different from any of the other tissues analyzed.

The ovary pathway correlation heatmap, however, shows a strong checkerboard patterning. This suggests that there are more numerous groups of genes that work tightly in sync with each other; yet, these genes oppose other tight groups of genes which is shown through the off diagonal regions of negative correlation. This strongly alludes to this modular regulation pattern, where some genes are tightly regulated with each other, and other genes are not. In addition, this also supports the notion that by looking at the gene-gene pair correlation within a pathway, we can potentially extrapolate sub-pathways that are activated in this specific tissue.


Normal B-cell Ribosome Pathway


Normal Brain Ribosome Pathway


Normal Breast Ribosome Pathway


Normal Skin Ribosome Pathway

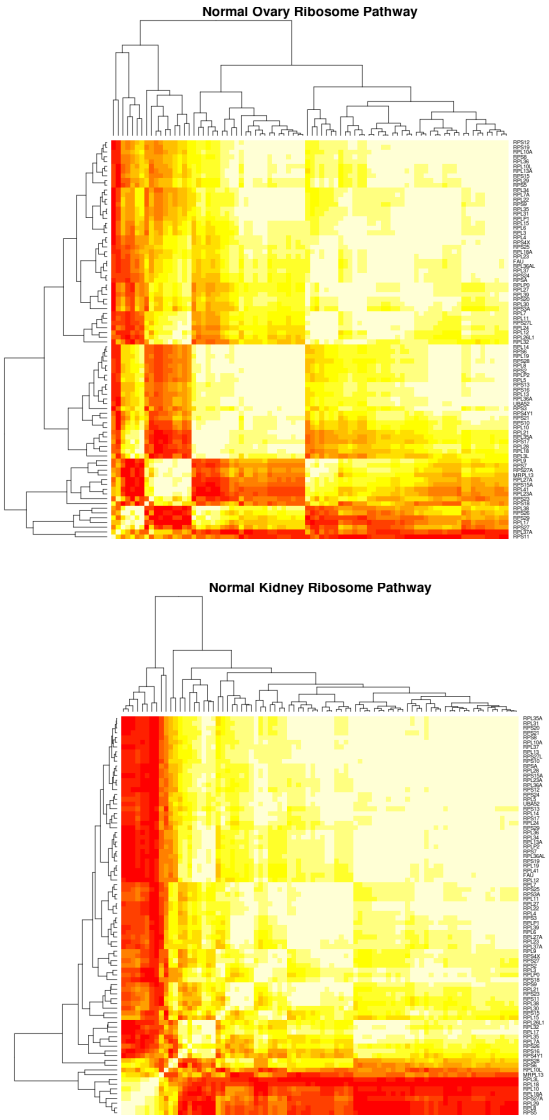


Figure 11. Heatmap of pathway correlation profiles for the Ribosome Pathway (hsa03010) in humans for each of the normal tissue data sets.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows and columns represent genes. (Yellow: positive correlations; red: negative correlations.) **(a)** Heatmap for normal B-cell. **(b)** Heatmap for normal brain. **(c)** Heatmap for normal breast. **(d)** Heatmap for normal skin. **(e)** Heatmap for normal ovary. **(f)** Heatmap for normal kidney.

In contrast, the skin, b-cell, brain and breast gene-gene pair correlation heatmaps show plaid patterns, where there are no strong groupings of genes within the profiles. This could be due to several reasons. These data sets had more samples, and as a result, the expected correlation values tended to decrease in magnitude due to an increase in variation of the gene's expression. Because of these more muted correlation values, identifying a strong signal from the pattern is more challenging. Another possible reason for this lapse in a strong modular signal from these tissue types is that there is not a strong regulative control over this pathway in these tissues. The Ribosome is essential for cell survival, but the extent of the coordinated expression of these genes to ensure cell survival is unclear. Assuming the Ribosome is a non-essential pathway for cell survival, i.e. lacking a strong regulative control, a strongly regulated correlation pattern would not be expected.

Clustering the genes based on gene-gene pair correlations for a given tissue type allows for identification of tissue specific correlation patterns. This shows which genes are working in tandem, and which genes are working against each other (i.e., repressive) for that tissue. However, this does not allow for a more global understanding of the pathway, with regards to multiple conditions or tissues. Alternatively to overcome this limitation, we can cluster the gene-gene pairs for each tissue type. From this approach, we now allow for the identification of pathway specific modules of regulation, as opposed to tissue specific modules of regulation. Via this approach, we can now see changes in the regulation of a pathway for different tissue types and potentially better understand how the pathway regulates itself.

Figure 12 shows the gene-gene pairs clustered for each of the six tissue types. Most notably, the kidney and ovary gene-gene pair correlations have a distinct profile, and in three clusters these two tissue types have opposite correlation values (i.e., when kidney has positive gene-gene pair correlations, the ovary has negative gene-gene pair correlations). A contradictory gene-gene pair correlation structure as that from the kidney and ovary could suggest tissue specific genes are involved in the ribosome. Knowing that previous studies have identified condition or developmental differences in expression of ribosomal genes [70], it can be interpolated that the same arises for different tissues, too.

There are still tissue specific correlation biases in the other 4 tissues, though not as prevalent likely due to reasons previously noted. Interestingly, the gene-gene pair correlation heatmap of the brain tissue in **Figure 11** shows two clusters of genes with correlated expression values. The same tissue type shows no distinct modules of regulation when viewed in **Figure 12**. In contrast, the skin tissue heatmap for the gene-gene pair correlations **Figure 11** does not show strong correlation clusters, but in **Figure 12**, modules of regulation are revealed. By clustering the gene-gene pair correlations across all the tissue types, we can gain information from one tissue type about its regulation structure, and then infer that structure onto other tissue types that might not have as clear of a pattern for regulation. Overall, these tissue specific correlation profiles suggest that the pathway has a modular pattern of regulation, and as such, certain attributes of the pathway work in a coordinated manner in one tissue and an alternate fashion in other tissues. Further, by clustering the gene-gene pair correlations, as completed in **Figure 12**, we now have the ability for extracting potential sub-pathways within the Ribosome pathway itself.

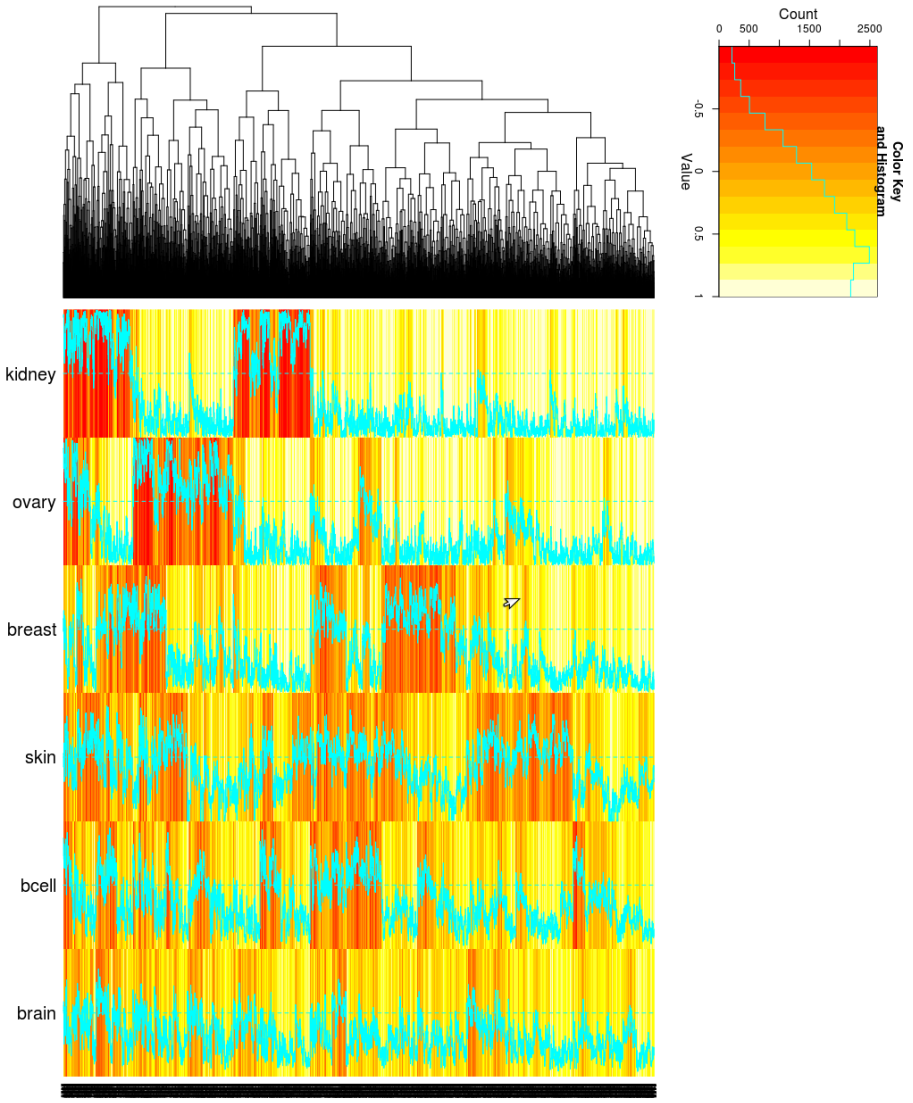


Figure 12. Heatmap of pathway correlation profiles for the Ribosome Pathway (hsa03010) in humans for each of the normal tissue data sets.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows represent normal tissue type and columns represent gene-gene pairs. (Yellow: positive correlations; red: negative correlations, cyan: actual correlation value.)

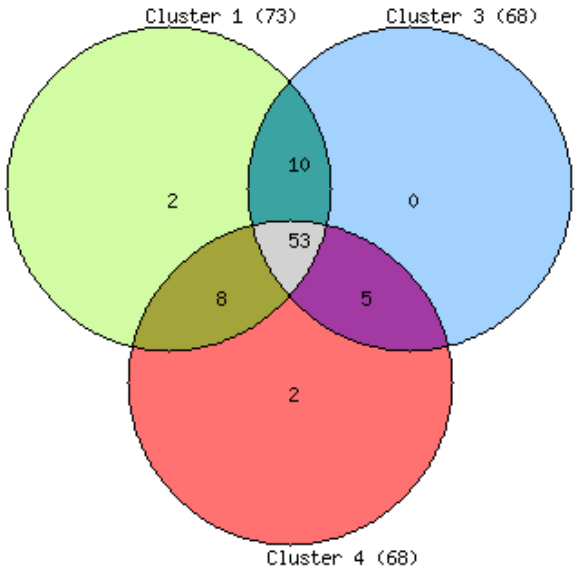


Figure 13. Venn diagram for select clusters of gene-gene pairs from the Ribosome Pathway (hsa03010) in humans.

Venn diagram showing the number of genes present in each of three select clusters from the clustering of gene-gene pairs (See **Figure 12**).

By clustering the gene-gene pair correlations across tissues, as opposed to clustering the gene expression values across tissues, we no longer force a gene into one cluster. This allows more flexiblity when trying to explain the dynamics of a pathway, besides the notion of being more consistent with biology. For example, it is possible for one gene to be involved in two different parts of a pathway or a member of two different protein complexes. When this gene is forced into one cluster, it no longer can be associated with all of its interaction partners, thereby missing potential and critical information. When we cluster with the pathway correlation profiles instead of the gene expression profiles, we can now assign a gene into a more proper number of clusters, and further identify all potential interaction partners from the data.

From the analysis done in **Figure 12**, we were able to identify three clusters of gene-gene pairs that have differential correlation patterns, especially in reference to the kidney and ovary tissue types. For each of these clusters, we isolated the set of genes that is inolved in the cluster of gene-gene pairs and found the overlap of genes within each of these three clusters (**Figure 13**). A majority of the genes from these three clusters (53 total) were present in each cluster, respectively. When considering the biology, these 53 genes present in each cluster are possibly the core proteins that constitute the ribosome complex, and the genes exclusive to particular clusters potentially could be the genes responsible for the tissue specific expression of the ribosomal proteins.

However, more interesting are the genes that are present in just one of these clusters. From Cluster 1, RPS18 and RPS20 are present in this cluster and none of the other two clusters while RPL18A and RPS24 are present in only Cluster 4. These genes are likely to be the driver genes that establish these distinct pathway correlation profiles for each tissue and the 53 other genes that are present. An alternative view of this is that there is minimal intra-correlation between these 53 genes, and strong inter correlation between each of these 53 genes and those present in just one or two of the clusters.

Understanding the pathway dynamics and regulation for a complex species like humans is difficult. Many pathways, especially the Ribosome Pathway, are present in simpler organisims, such as bacteria. It is in these species that we can really start to understand the modular design of pathways and how this relates to their regulation. In bacteria, many genes are organised in the genome by operons, or sets of genes that are transcribed in one transcript and subsequently processed into their individual genes post-transcriptionally. If a set of genes are truly part of the same module, then it is likely that these genes are part

of the same operon in the more primitive species. To determine the importance and revelance, we have applied the same approach to the *E. coli* pH data set previously used.

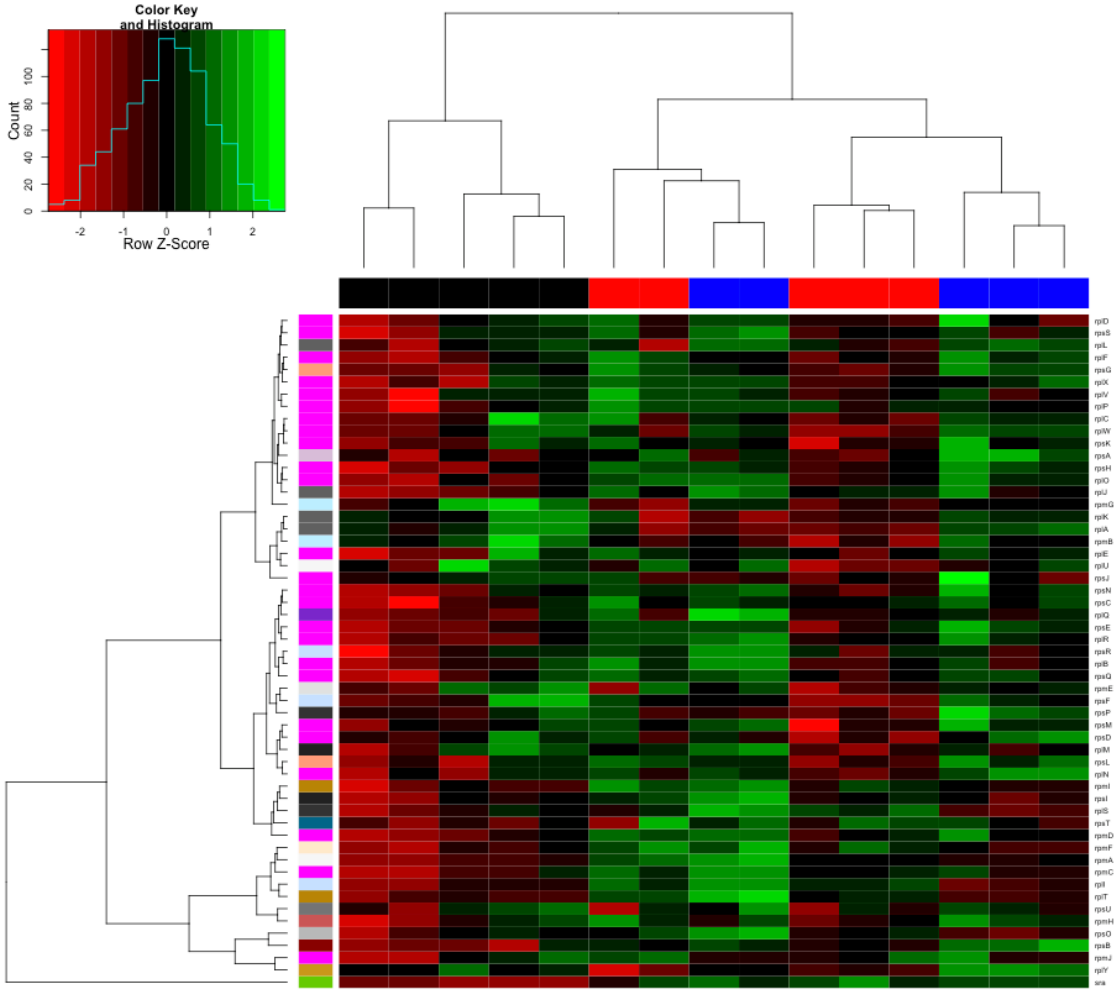
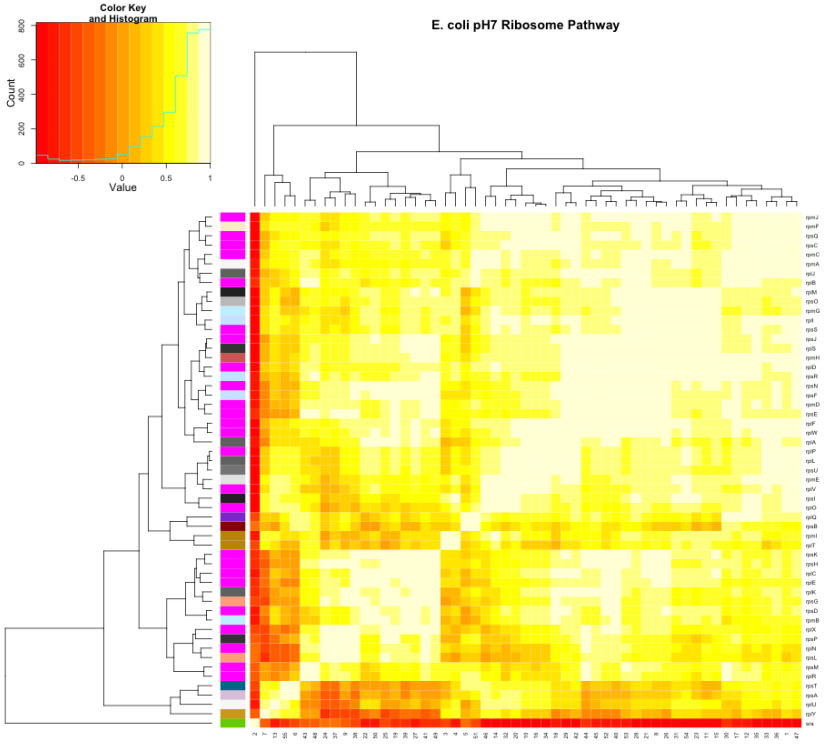
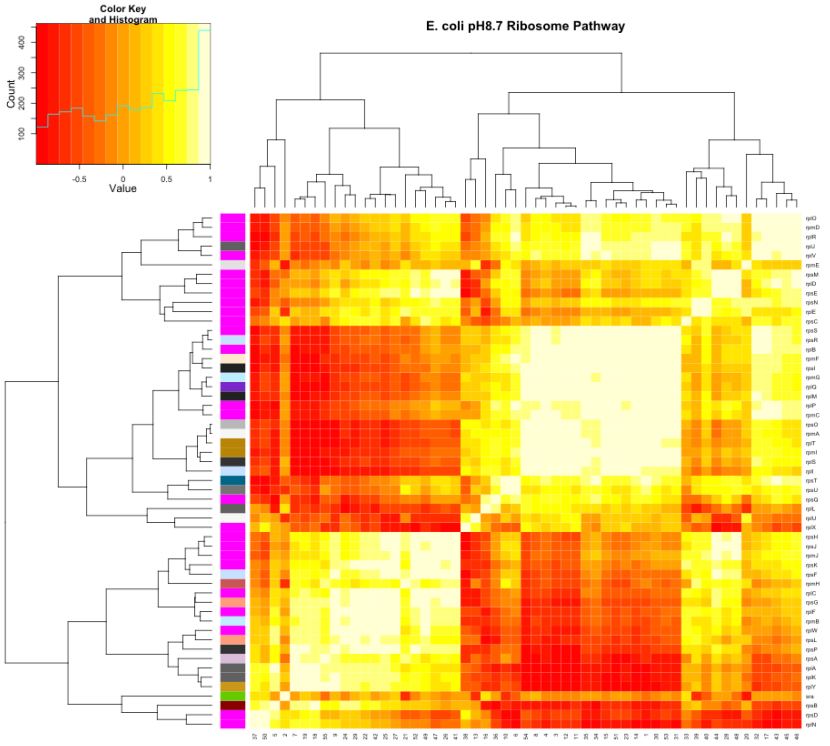


Figure 14. Heatmap of the gene expression for those genes in the Ribosome Pathway (ecj03010) in E. coli for three pH conditions, respectively.

Heatmap and clustering of genes are based on gene expression signal. Rows are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; Red: negative z-normalized expression level) Row color bar: Black—pH7, Blue—pH8.7, Red—pH5) Column color bar: identify operons.


Color Key
and Histogram
Count
Value
E. coli pH7 Ribosome Pathway


Color Key
and Histogram
Count
Value
E. coli pH8.7 Ribosome Pathway

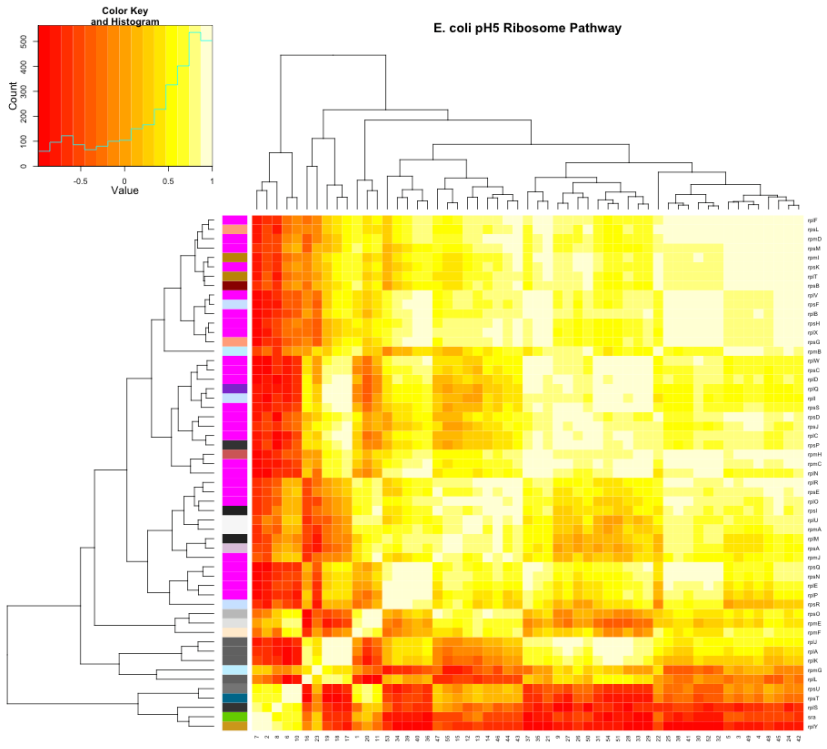


Figure 15. Heatmap of pathway correlation profiles for the Ribosome Pathway (ecj03010) in E. coli for each of the three pH conditions, respectively.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows and columns represent genes. (Yellow: positive correlations; red: negative correlations; Row bar represents operon.) **(a)** Heatmap for pH7. **(b)** Heatmap for pH8.7. **(c)** Heatmap for pH5.

The Ribosome Pathway in *E. coli* is tightly regulated to make only enough of each protein that it needs and have no excess. To accomplish this, many of the ribosomal genes are clustered into large operons, or sets of genes that are transcribed in one transcript. By being in the same regulation cluster, the intra-correlation of these genes should be high, when compared to genes that are not apart of this cluster. Moreover, these

gene clusters and operons should be a first insight into these modules of regulation from an evolutionary perspective.

We classified each gene into an operon based on its relative location and direction of transcription in the *E. coli* genome. With this additional data, we plotted a heatmap of the gene expression levels for each of the Ribosomal genes under the three separate pH samples (**Figure 14**). From this, we can see that those samples from pH7 (denoted by black) cluster together with similar expression profiles and the samples from the pH8.7 (blue) and pH5 (red) cluster within each other. This suggests that the Ribosome pathway has an expression pattern for normal conditions, and an alternate expression pattern for abnormal environmental stimuli, in this case a change in pH. In the case of the Ribosome Pathway, the type of stimuli seems not to have a huge effect on the expression profile, rather the presence of the stimuli is more important.

When clustering the genes based on their expression profiles for each of the conditions, many of the genes from one of the largest operons (denoted in magenta) clustered together. This substantiates that the inclusion in an operon has a correspondence with the levels of gene expression. One gene, the SRA gene, did not cluster with any of the other genes. This gene, the Stationary-phase-induced ribosome-associated protein, is expressed only when the cells are in stationary phase, i.e., not replicating. The SRA gene is one of the few that are universally down-regulated in the normal pH7 and up-regulated under the stressed pHs. During these alternate pH conditions, the cell's primary goal is survival and not replication. As such, the *E. coli* cells will remain in stationary phase, thereby expressing the SRA gene at greater levels.

Gene expression alone does not give us a good understanding about the dynamics of pathway regulation. A second dimensional analysis using the pathway correlation analysis is shown in **Figure 15**. Here, the pathway correlation profile for pH7 showed strong positive correlations with everything except the SRA gene. This once again confirms that the SRA gene has a negative correlation with all other genes under normal conditions, since these cells are not in stationary phase. At pH8.7, three distinct partitions arise, two of which have strong inter-partition negative correlations. This suggests that there are two groups of genes that work together, but have negative influences on each other. Interestingly, there is no strong operon bias towards these clusters in any of the three pH conditions. A possible reason that the correlation profiles do not align well with the operon profiles could be due to post-transcriptional processing rates, degradation rates as well as functional usage rates, especially in regards to the rRNA genes.

Considering a third dimensional perspective, clustering on the gene-gene pair correlations for each of the three pHs, we can identify clusters of differential correlation patterns between each of the conditions (**Figure 16**). At pH7 there is not a strong negative correlation presence, likely due to the SRA gene. However, pH8.7 and pH5 have complimentary correlation patterns in several prominent partitions. This alternating correlation pattern, not seen at the gene expression level, suggests that there is a global Ribosomal response under stressing pHs, and each pH has its own, independent regulation pattern. This also confirms modules of regulation that were not present from a gene expression analysis.

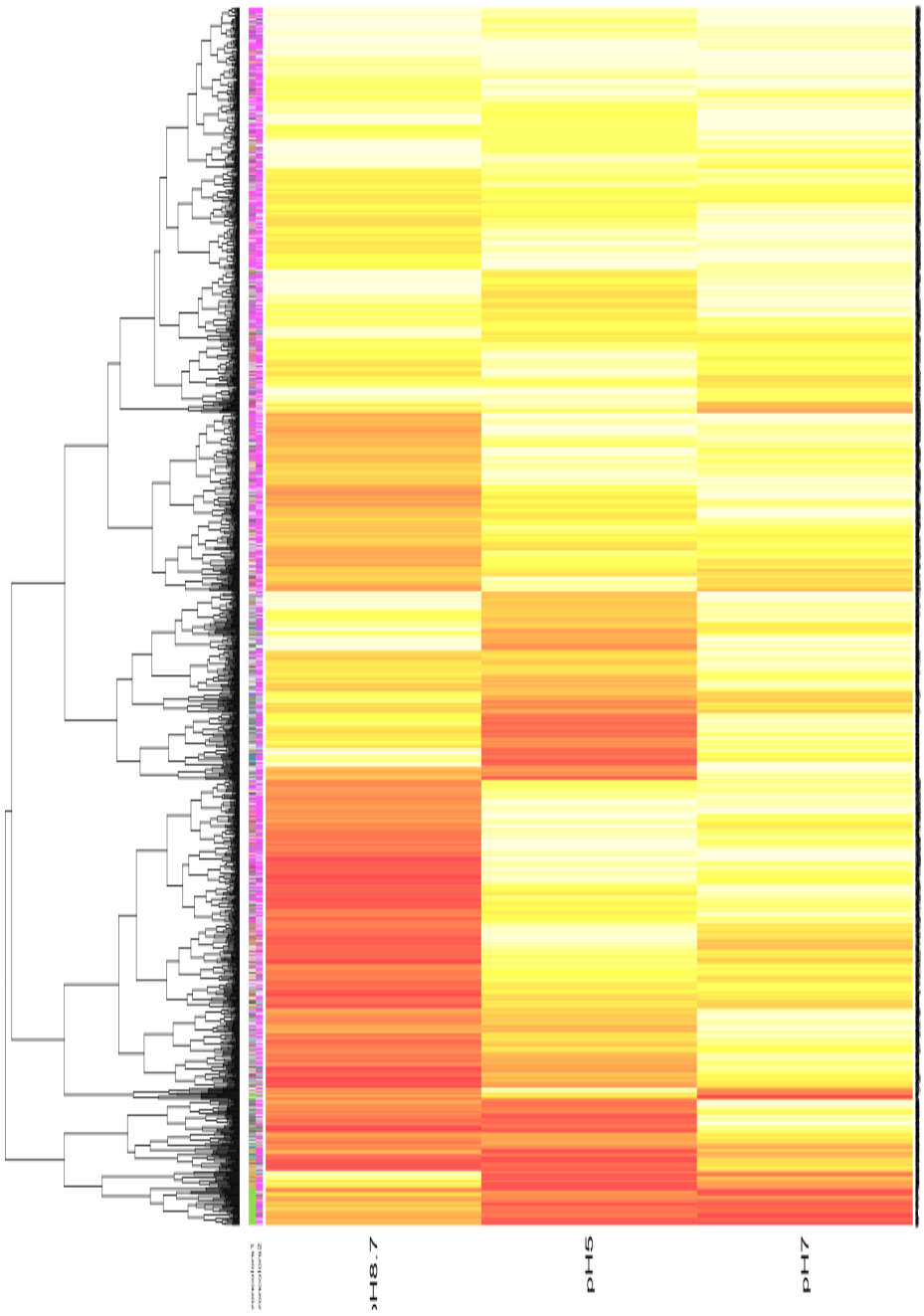


Figure 16. Heatmap of pathway correlation profiles for the Ribosome Pathway (ecj03010) in E. coli for each of the three pH conditions, respectively.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Columns represent normal tissue type, rows represent gene-gene pairs, and row bar represents operon. (Yellow: positive correlations; red: negative correlations)

4.4 Conclusion

Through the pathway correlation profile analysis, we can now better characterize the pathway dynamics of the Ribosome pathway (03010). From our novel clustering approach, we no longer force one gene into one cluster, thereby establishing a clustering protocol that is more in line with the biology of pathway dynamics. Furthermore, we can now identify module specific genes that are most likely responsible for the different gene-gene pair correlation profiles, and therefore the regulation of that module.

We further investigated the use of operons as a justification for the modules of regulation, i.e., sub-pathways. Though initially, there was no strong correspondence between operons and groups of genes working together, we cannot all together exclude the notion that operons are an initial framework for modules of regulation. A cross species comparison could be used to help confirm the connection between operons and modules of regulation, especially when comparing prokaryotes and eukaryotes. Preliminarily, when comparing sequence alone, alignments of select human ribosomal genes against the *E. coli* genome produce some strong sequence alignments. These toy alignments, however, did not align human ribosomal genes to *E. coli* ribosomal genes, rather to predicted genes. Nonetheless, showing a sequence conservation in few genes opens the door for an analysis that investigates the conservation of ribosomal function and regulation throughout the evolution of species.

The additional uses shown here of the pathway correlation profiles, besides identifying and ranking pathways based on the significance of their perturbations, allow for more comprehensive studies of the biological context. The added perspectives that the approach gains have the potential for many important and relevant applications. With

more systematic approaches to studying pathways, i.e., a meta-analysis type approach, global understandings of pathways can result, as opposed to condition or disease specific views.

5 Chapter 5: Applications to colorectal cancer

5.1 Introduction

Understanding cancer as a disease and its progression is not trivial. In many instances, there is not one mishap, but a calamity of errors that ultimately results in cancer [71]. Moreover, as cancer progresses each stage takes on its own phenotype, especially as demonstrated by the trend towards personalized medicine [72]. In some instances, the stage can be determined morphologically, but more relevant is genotypic staging, since in the end it is the genotypic and expression behavior that dictate treatment, prognosis and outcomes. Due to the importance of genotypic staging, developing computational protocols that can better interpret the gene expression data already available is essential. Here, we complete a case example using colorectal data and focusing on the MAPK Pathway.

The mitogen-activated protein kinase (MAPK) Pathway (hsa04010) is a signal transduction pathway that plays an important role in cell proliferation, differentiation, apoptosis, angiogenesis, and metastases. Due to its involvement in such vital cellular processes, it is essential that the entire system is in balance in order to prevent oncogenesis. In addition to being involved in these processes, the MAPK pathway also contains several known onco- and proto-oncogenes, including but not limited to c-myc, c-fos, c-jun, Raf, and Ras [73,74]. The MAPK Pathway has been linked to colorectal cancer and therefore studied more extensively[73,75,76]. Due to this linkage between the MAPK pathway and cancer, we will focus our study on trying to identify regulative aberrations of the MAPK Pathway with respect to colorectal cancer.

We believe that as cancers progress from early stages to later stages, pathway regulation evolves and rewires. In addition, it has been shown previously that different cancers rewire pathways in different ways throughout oncogenesis [77]. Here, we utilize the pathway correlation profile method as a tool to try to characterize and understand the Ribosome and MAPK pathways through the progression of colorectal cancer from normal tissue, to polyps, through the four stages and finally ending with distant metastasis. From this approach, we can look into the dynamics of these two pathways for normal tissues, and see how and where the pathway correlations change during the disease progression. Lastly, we will end by showing how pathway correlation profiles can be used as a potential tool for identifying disease related network biomarkers.

5.2 Data

When studying disease progression, it is essential to use a data set with distinct disease states during the progression from normal to advanced disease. To accomplish this, we have used colorectal cancer for this case study [78]. Colorectal cancer has distinct stages morphologically [79] and large genomic changes throughout tumorigenesis [78,80]. Due to the distinct classification of patients, and the progressive changes throughout the disease progression, this data set is an ideal choice for a systematic pathway regulative dynamics analysis.

Here, we have used a colorectal data set downloaded from GEO database (GSE41258) [78]. This study thoroughly investigates the progression of colorectal cancer, including normal colon tissue, samples from polyps, stage I, stage II, stage III, stage IV, metastasis to the liver and lungs, as well as normal liver and lung tissue for comparison. In total, the data set has 390 samples, with disease stage specific totals in Table 3. The cell line and

select other arrays were not used in this analysis. Once the gene expression data was obtained, RMA was used on each condition for normalization and background correction, and the pathway correlation method was used for each condition to make profiles for the Ribosome (hsa03010) and MAPK (hsa04010) Pathways as seen in **Figure 17** and **Figure 18** respectively.

Table 3. Allocation of samples for colorectal cancer data set.

Condition	No. of Samples
Normal Colon	54
Normal Liver	13
Normal Lung	7
Polyp	48
Stage I	28
Stage II	48
Stage III	49
Stage IV	57
Liver Metastasis	47
Lung Metastasis	20
Cell Lines	12
Others	7

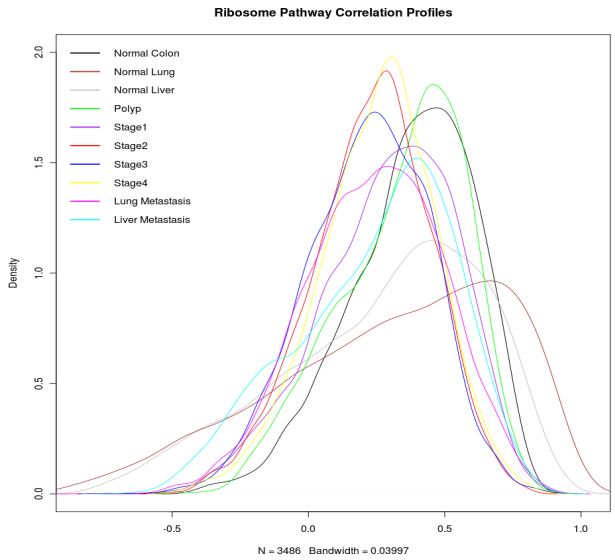


Figure 17. Pathway correlation profiles for Ribosome Pathway (hsa03010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

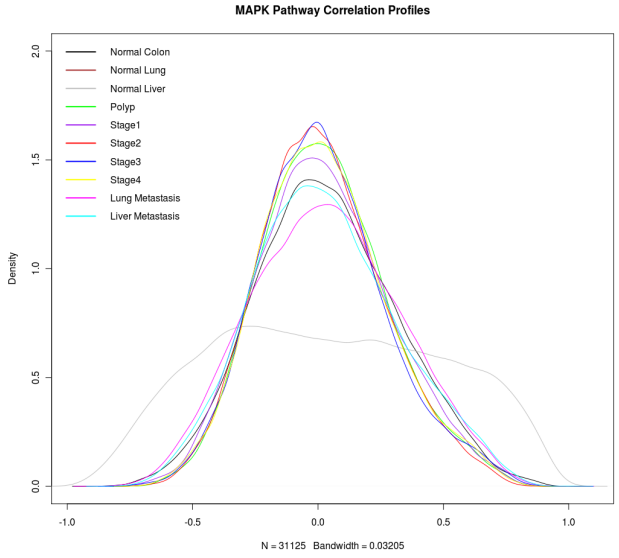


Figure 18. Pathway correlation profiles for MAPK Pathway (hsa04010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

5.3 Results and Discussion

Completing microarray and gene expression analysis on cancer data sets is difficult due to the limitations on number of samples, increased variance within data types, combinatorial amounts of abnormalities that could cause the disease, and even incorrect classification of disease state. By reducing the number of variables, e.g., looking at a single pathway at a time, we can attempt to overcome these challenges in a systematic way. For this dissertation, I look into the regulative pathway dynamics of the Ribosome (hsa03010) and the MAPK (hsa04010) Pathways and how these dynamics change throughout the tumorigenesis of colorectal cancer.

The Ribosome is essential for translation activities, and increased Ribosomal activity is expected in cancer cells due to increased cellular actions and differentiations [81]. **Figure 19** shows the relative gene expression levels for those genes in the Ribosome Pathway across all conditioned samples. Many of the normal colon, liver and lung samples show decreases in expression of the Ribosome genes, relative to the various stages of colorectal cancer. Alternatively, the polyp and four stages of colorectal cancer show universal increases in gene expression of the Ribosomal genes, relative to the normal tissue samples. Interestingly, however, numerous samples from the lung and liver metastasis samples show decreases in gene expression of this pathway. This observation is not consistent with the expectation for increased Ribosomal activity in all cancer cells. However, these results do suggest that there could be regulative changes during the process of metastasis, which result in less of a need for ribosomal expression.

When considering the sample clustering completed in **Figure 19**, strong clusters of one condition type are not present, with the exception of two disjoint normal colon, normal

liver and liver metastasis clusters. This lack of strong stage and condition clustering suggests that the gene expression of the Ribosome Pathway does not differentiate the samples successfully, thereby not necessarily being a good predictor of staging for colorectal cancer.

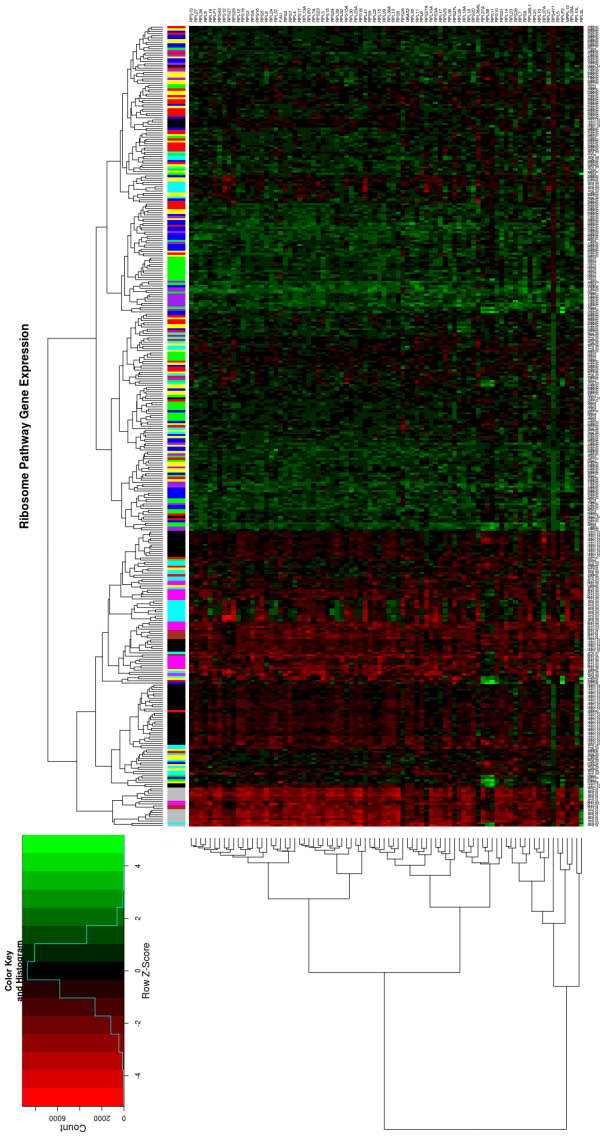


Figure 19. Heatmap of the gene expression for those genes in the Ribosome Pathway (hsa03010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Heatmap and clustering of genes are based on gene expression signal. Columns are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; Red: negative z-normalized expression level) Row color bar: identify tissue type.

The gene expression profiles for the genes in the Ribosome Pathway had minimal information, and were not successful in differentiating out the various stages of the cancer. By using the pathway correlation profiles for each stage, instead of the gene expression profile, then clustering based on the gene-gene pair correlations, we can use the data in a new and novel way (**Figure 20**). First, we look at the disease progression clustering of the columns in **Figure 20**, where we can see that all four stages of colorectal cancer cluster together, followed by the lung and liver metastasis. Metastasized cancer samples result after cells from the original tumor migrate to a new location, thereby maintaining many of the same genomic properties of the primary tumor, and not the same properties as the location of the metastasis' normal tissue. This notion is supported in this data set; after clustering based on gene-gene correlations, the two metastatic samples clustered with the primary tumor samples. In addition, these two samples do not cluster with their respective normal tissue at the site of metastasis, i.e., lung and liver, further supporting the fact that metastatic samples are cells with genetic properties of the original tumor.

Moreover, the hierarchical clustering shows that the normal colon and polyp samples' gene-gene pair correlation profiles are most closely related. Studies have suggested that polyps are a precursor to colorectal cancer, without being cancer themselves. Having normal colon and polyp conditions cluster together show that the polyp resembles a

regulation profile closer to that of normal than to the cancer stages, thus supporting that the polyp is not cancerous and substantiating the colorectal cancer pre-cursor suggestions.

Considering the clustering of the gene-gene pair correlation profiles across all the conditions, a majority of the gene-gene pairs have positive correlations, which has been seen numerous times for the Ribosome Pathway genes. This is likely due to the fact that this pathway encodes for a complex, in which the majority of the genes need to be expressed at similar levels, thus resulting in positive gene-gene pair correlations. Moreover, there are not as distinct of partitions as seen in the *E. coli* and human normal tissue samples from the previous chapter. Overall, the cancer samples have gene-gene pair correlations that are lower in magnitude. This muted nature of the pathway correlation profile for each stage is likely due to the increased variance of the patient samples, which is frequently seen in oncogenic data sets. However, several conclusions can still be made. The gene-gene pair correlation profiles of the normal lung and liver tissues are distinct, and in several clusters, there are opposite correlation profiles for the same gene-gene pairs, which is in accordance with previous results that suggest tissue specific regulation of pathways. The colon tissue derived profiles (i.e., normal colon, polyp, stages I-IV, and metastases) show similar correlation profiles, although each condition has its own distinct feature.

Next, we divided the data into three partitions and isolated the genes involved in each partition so we can identify potential gene specific partitions (**Figure 21**). There were a total of 76 genes in the Ribosome Pathway, and 50 of these genes were present in each of the three partitions. Seeing 50 genes present in each of the partitions again suggests these

50 genes are more likely to be members of the core set of proteins for the ribosomal complex, i.e., expressed and present under all tissue types and conditions.

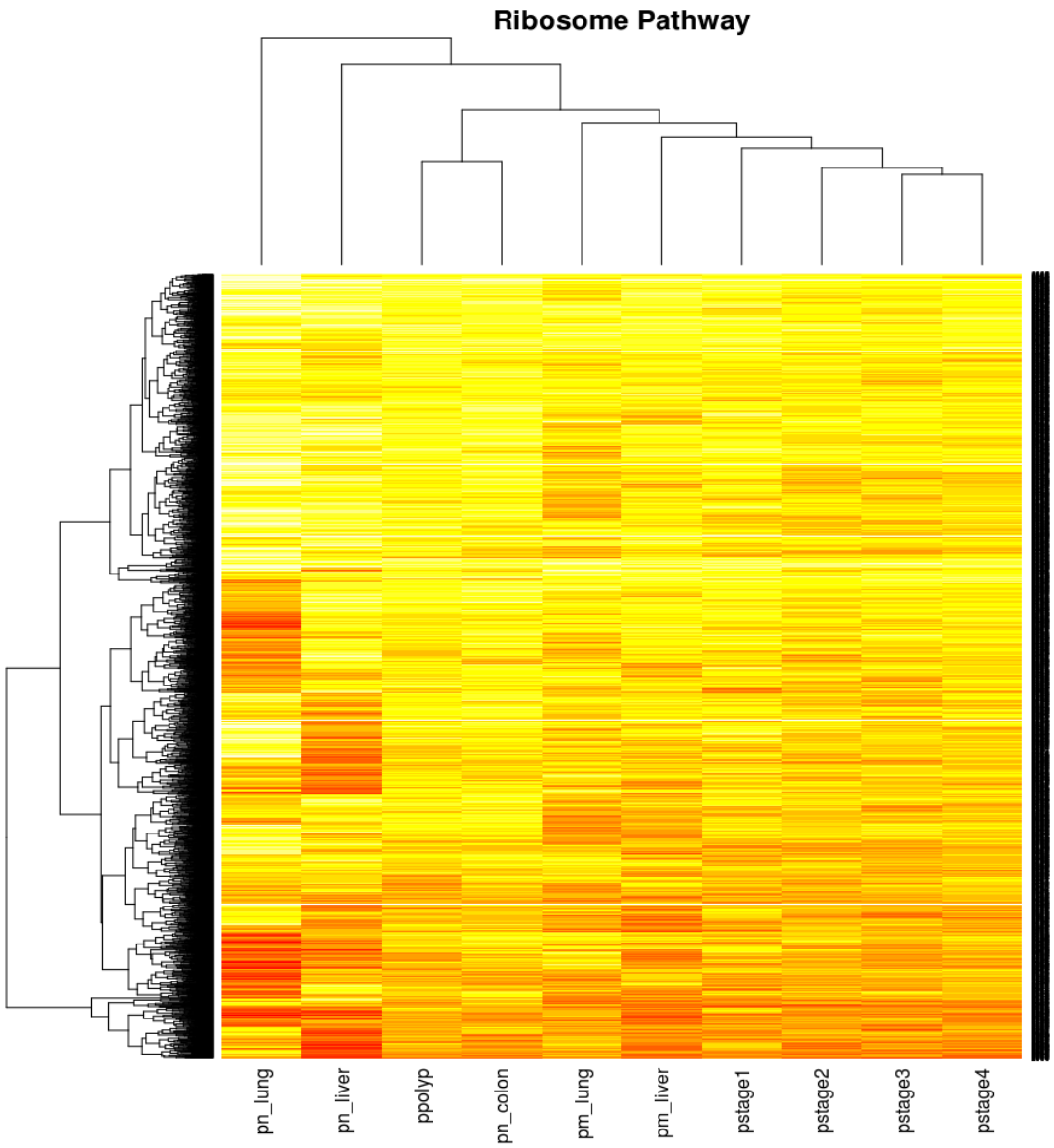


Figure 20. Heatmap of pathway correlation profiles for the Ribosome Pathway (hsa03010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows represent gene-gene pairs and columns represent patient tissue types (from left: normal

lung, normal liver, polyp, normal colon, metastasis to lung, metastasis to liver, stage 1, stage 2, stage 3, stage 4). (Yellow: positive correlations; red: negative correlations, cyan: actual correlation value.)

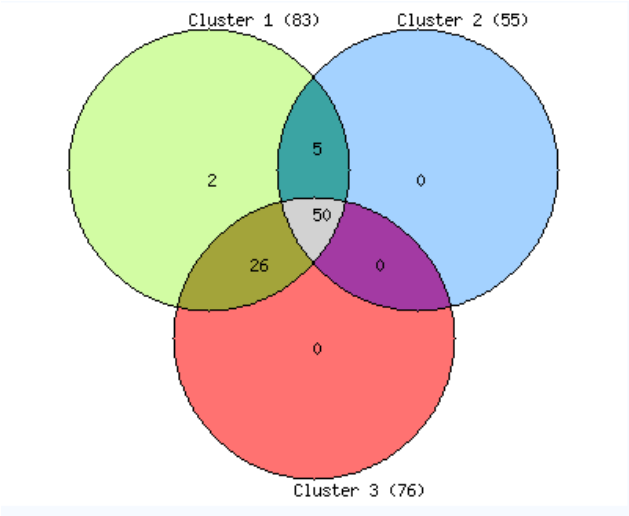


Figure 21. Venn diagram for select clusters of gene-gene pairs from the Ribosome Pathway (hsa03010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Venn diagram showing the number of genes present in each of three select clusters from the clustering of gene-gene pairs (See **Figure 20**).

Two genes, however, were only present in one partition: ribosomal protein S21 (RPS21) and ubiquitin A-52 residue ribosomal protein fusion product 1 (UBA52). Ubiquitin is essential in directing proteins to the proteosome for destruction [51]. Abnormal ubiquitin signaling has been linked to several cancers, including colorectal cancer [82]. Due to this link with oncogenesis, ubiquitin and the ubiquitin networks have also been considered a target for many cancer therapy studies [83,84,85]. By identifying the ubiquitin related

gene as a member of just one cluster, this demonstrates that there is a tissue/condition specific regulation of this gene, with respect to the entire ribosomal complex, in colorectal cancer. As a result, this again could be a potentially targeted mechanism when considering alternative therapeutics for the treatment of colon cancer.

Understanding the Ribosome Pathway is important because of its relatation to post transcriptional regulation and protein translation, both necessary steps for cellular survival. However, the MAPK pathway has also been linked to colorectal cancer and therefore studied more extensively [73,75,76]. Moreover, the MAPK Pathway is a signal transduction pathway known to be perturbed in many cancer types. By being a signal transduction pathway, as opposed to a protein complex driven pathway like the Ribosome, we have an added dimension when considering the biological implications of the results obtained. By using the pathway correlation profile as a tool to analyze the MAPK pathway, we can potentially identify alterations and abberations in the signaling of this pathway, thereby offering an additional perspective on the data and results.

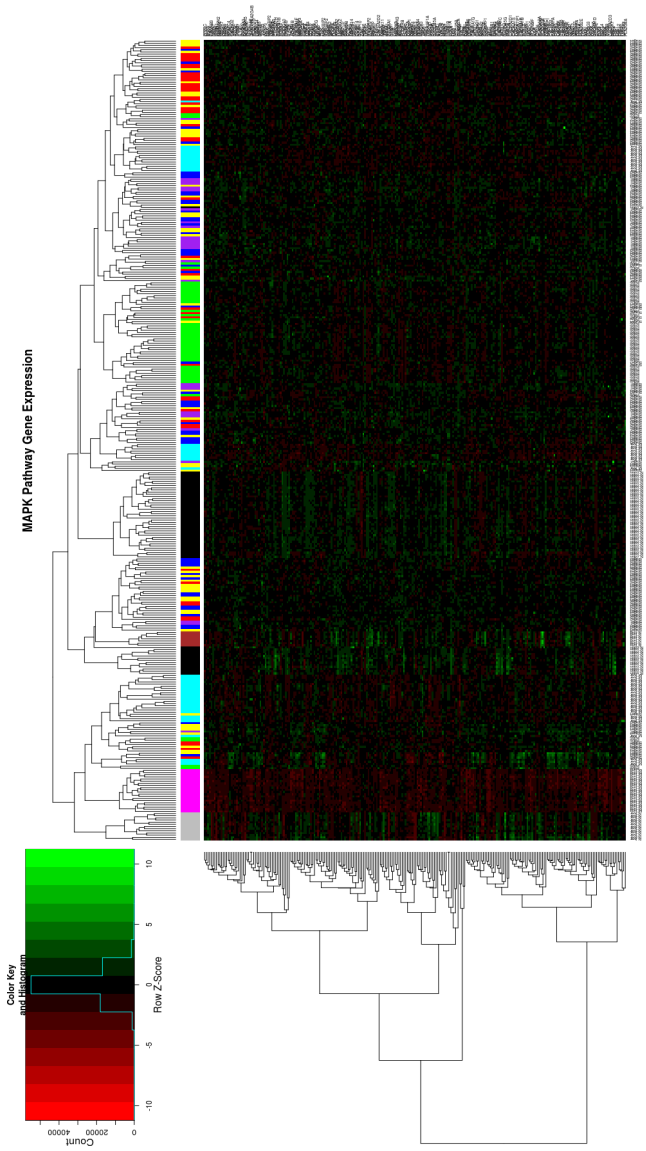


Figure 22. Heatmap of the gene expression for those genes in the MAPK Pathway (hsa04010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Heatmap and clustering of genes are based on gene expression signal. Columns are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; Red: negative z-normalized expression level) Row color bar: identify tissue type.

Traditional gene expression analysis considers individual gene expression values for each of the samples. **Figure 22** shows the gene expression for those genes in the MAPK pathway and for each sample in the colorectal cancer data set. The samples were clustered via hierarchical clustering based on their gene expression profiles as well as the genes. Each gene was z-normalized, whereby a green cell means that sample has higher expression levels of the gene compared to the rest of the samples, and the converse for red cells. Overall, there are not extreme changes in gene expression for those genes involved in the MAPK pathway. Even so, gene expression levels alone do not infer dysregulation of a pathway.

However, more significantly is the clustering of the samples for the MAPK genes. Previously, when clustering the samples based on the genes from the Ribosome Pathway, there were few clear clustering of stages or tissue samples, with the normal colon being the most decisive. The MAPK Pathway shows strong groupings of normal colon (black), normal lung (brown), normal liver (grey), polyp (green), liver metastasis (teal), and lung metastasis (magenta). These tissue/condition related groupings suggest that the MAPK genes do have different expression profiles for each condition, and minimal differences within one data type. Additionally, the differential expression profile also encourages the use of the MAPK pathway as a discriminator of the stages of colorectal cancer.

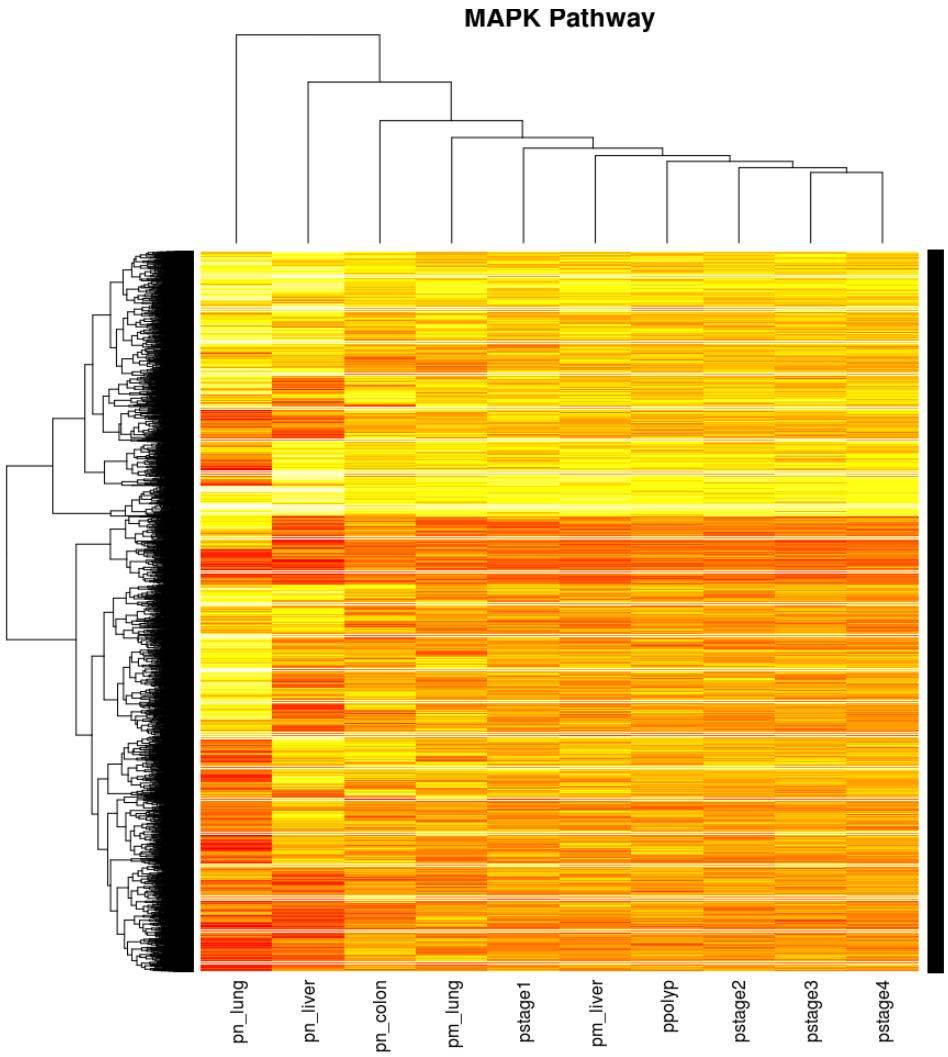


Figure 23. Heatmap of pathway correlation profiles for the MAPK Pathway (hsa04010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Heatmap and clustering of genes are based on their gene-gene pair correlations. Rows represent gene-gene pairs and columns represent patient tissue types (from left: normal lung, normal liver, polyp, normal colon, metastasis to lung, metastasis to liver, stage 1, stage 2, stage 3, stage 4). (Yellow: positive correlations; red: negative correlations, cyan: actual correlation value.)

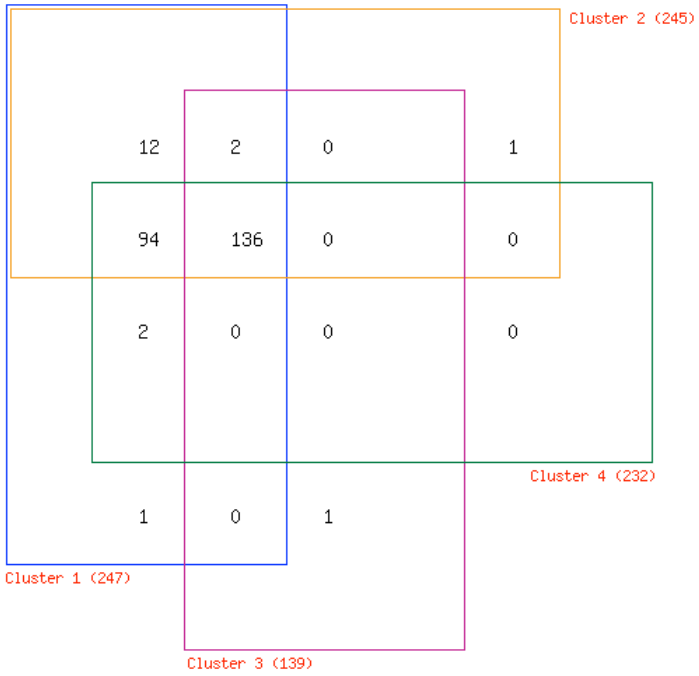


Figure 24. Venn diagram for select clusters of gene-gene pairs from the MAPK Pathway (hsa04010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Venn diagram showing the number of genes present in each of three select clusters from the clustering of gene-gene pairs (See **Figure 23**).

An alternative perspective on the MAPK Pathway considers the gene-gene correlations, and how these correlations change throughout the tumorigenesis. **Figure 23** shows the clustering for the MAPK pathway based on gene-gene pair correlations over all tissues and conditions. Like the clustering in **Figure 20**, the normal lung and liver tissues do not cluster with the result. However, the rest of the samples do not cluster as expected like in the Ribosome Pathway. For example, it is surprising that the polyp correlation profile cluster closer to stages II-IV, and not adjacent to the normal colon or even stage I. Again,

this confirms that under each stage, there exists a differential regulation profile for the MAPK pathway.

In regards to the clustering of the gene-gene pairs in **Figure 23**, there are some strong partitions that aid in the differentiation of the tissue and condition correlation profiles. To better understand the gene-gene pair clustering, we partitioned the gene-gene pairs into six subdivisions, and selected four clusters from these partitions. When we decompose the genes involved in each cluster, we can see the overlap of genes present in each of these four clusters (**Figure 24**). Three out of the four clusters each independently contain one gene: neurofibromin 1 (NF1); nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4 (NFATC4); and transforming growth factor, beta 1 (TGFB1). By being a strong influence in only one cluster, and not present in the others, these genes are likely to be strong regulative drivers for the MAPK Pathway at a specific disease stage time point. In fact, the NFAT family of transcription factors has been linked to several cancers and their progressions [86] and mutations in NF1 predisposes one to certain malignancies [87], due to its role in regulating RAS. There were also two sets of pairs that were present in only 2 clusters: MAP3K1 and MYC; FOS and SRF. Of these, both MYC and FOS are known oncogenes [73], MAP3K1 has been linked with breast cancer [88], and SRF is a transcription factor where increased expression has been seen in liver metastasis [89]. Recognizing these select genes that are strongly involved in just one or two clusters allows us to pinpoint potential genes that drive the abnormal signaling of the MAPK pathway at different stages of colorectal cancer. On top of that, seeing the involvement of these genes in other cancers buttresses our prediction that they are also involved in the progression of colorectal cancer.

If we look closer at the gene expression of these seven genes (**Figure 25**), we can find some interesting patterns. Firstly, when looking at SRF, we find decreased gene expression levels in the liver and lung metastasis, which is not consistent with what was reported by Choi et al [89]. A reason for this difference could have resulted because their experiment was assaying the protein expression, whereas this data is gene expression data. There is a marked decrease in expression of MYC for the normal samples, which is consistent given the oncogenic status of MYC. There is an increase in gene expression of NFATC4 in the later stages of colorectal cancer, but a sharp decrease in expression of this gene once metastasis has occurred. This increase in expression during the later stages is partially expected due to the role of NFAT with metastasis [90], however the decrease in expression once metastasis has occurred is not as expected. Lastly, NF1 shows decreased expression in the liver and lung metastasis samples. Taken together, these genes are again likely drivers of tumorigenesis and metastasis of colon cancer.

The last group of interest in **Figure 24** is a group of 12 genes that are only present in two of the four clusters: RPS6KA2, ELK4, TRAF6, MAPKAPK, PRKX, FGF13, H-RAS, PTPN7, MAP2K6, NTRK2, IL1B, TP53. Of these, the most notable are H-RAS and TP53, an oncogene and a tumor suppressor gene, respectively. If we consider their gene expression (**Figure 26**), we notice a general decrease in most all the gene's expression in the lung metastasis samples. Also, we see increases in the gene expression of NTRK2, a tyrosine kinase receptor, and IL1B, a signaling protein, in normal colon and lung samples, but not normal liver. Further, we notice a decrease in TP53 in lung metastasis, however no strong change in gene expression in most all other samples. Alterations to the expression of signaling proteins and tyrosine kinase receptors can alter the signaling of

the MAPK pathway, as well as changes in oncogene and tumor suppressor genes will reinforce the aberrant signaling of this pathway. Taken together, these genes have the potential to be additional regulators of the MAPK pathway with respect to colon cancer.

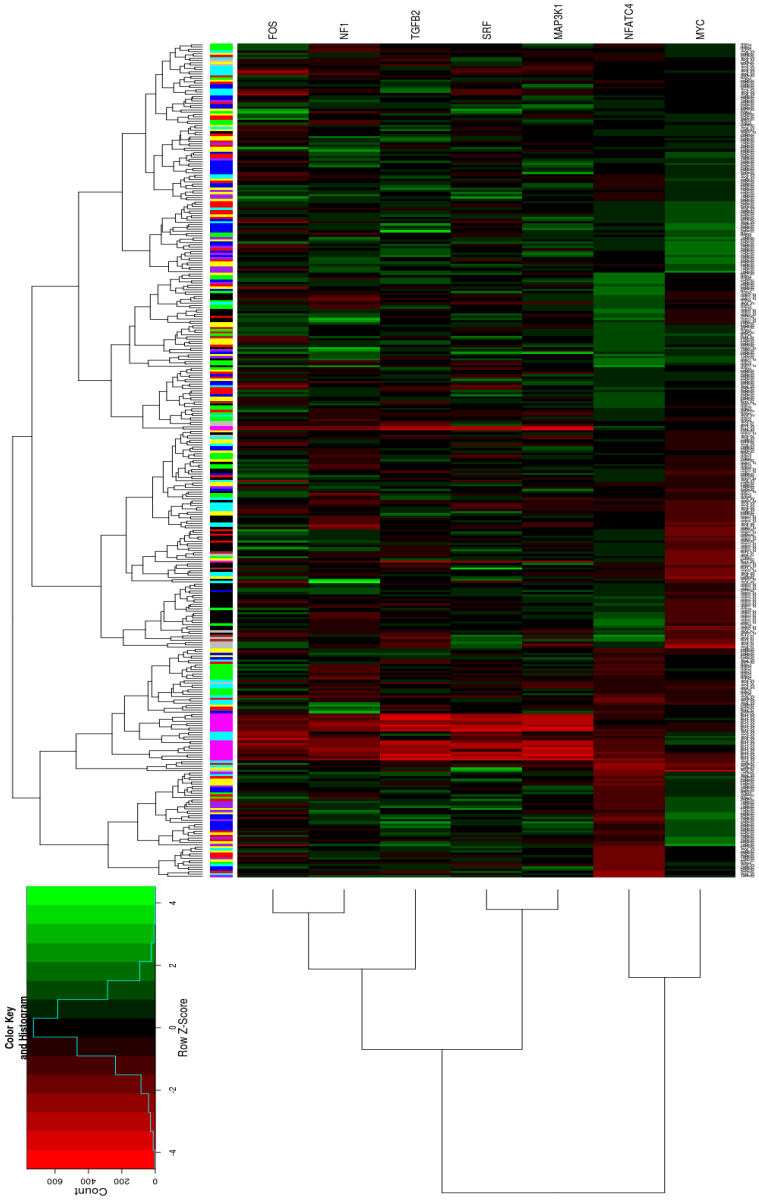


Figure 25. Heatmap of the gene expression for select genes in the MAPK Pathway (hsa04010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Heatmap and clustering of genes are based on gene expression signal. Columns are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; Red: negative z-normalized expression level) Row color bar: identify tissue type.

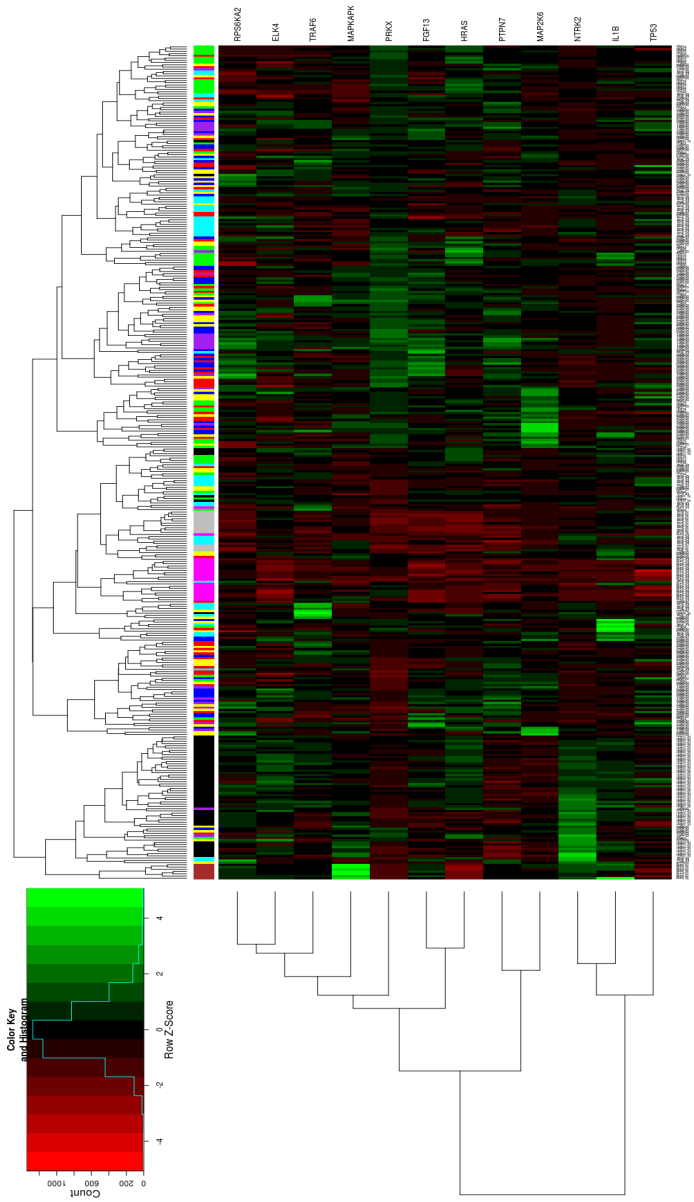


Figure 26. Heatmap of the gene expression for select genes in the MAPK Pathway (hsa04010) in humans for normal liver, lung and colon, polyp, stages I-IV of colorectal cancer, and metastasis to lung and liver data set.

Heatmap and clustering of genes are based on gene expression signal. Columns are normalized to show a more relative expression level compared to other samples. (Green: positive z-normalized expression level; Red: negative z-normalized expression level) Row color bar: identify tissue type.

Further, we can look into the connectivity of the genes within each partition of the gene-gene pair correlation clustering heatmap. When the hierarchical clustering of the gene-gene pairs is divided into six partitions, five out of the size partitions each contain just one connected component, when considering each gene as a node and each gene-gene pair correlation as an edge. The sixth partition, however, contains seven connected sub-components (Figure 27, Figure 28 and Figure 29).

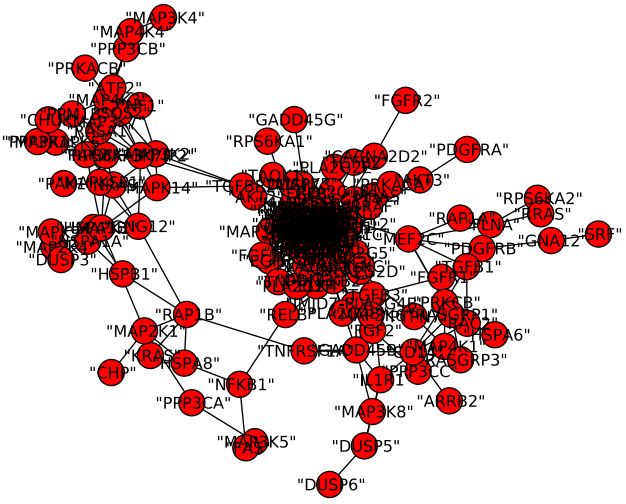


Figure 27. Network drawing of largest component from partition.

Each node represents a gene and each edge represents a gene-gene pair correlation.

Figure 27 shows the largest component from this cluster of gene-gene pairs. Overall, there is a large group of genes that are correlated with each other, possibly suggesting a set of essential genes for the MAPK pathway. Expanding from the main cluster in this sub-component, there are some auxiliary genes that are correlated with just one or two other genes in the network. These auxiliary genes are not likely to be strong drivers of the pathway, rather passenger genes for the pathway.

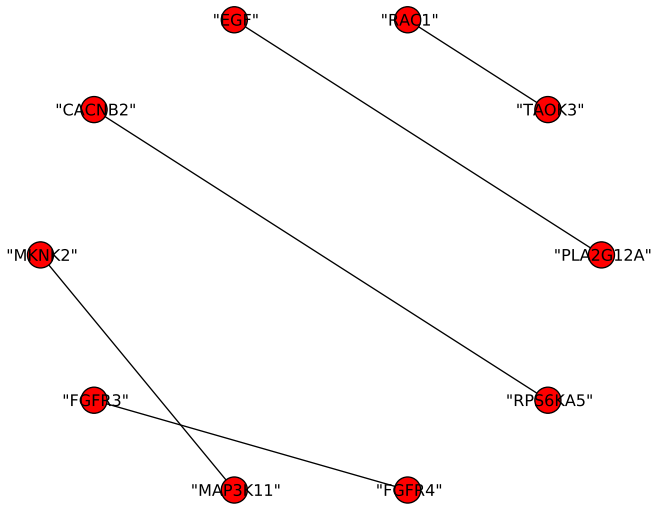


Figure 28. Network drawing of five component from partition.

Each node represents a gene and each edge represents a gene-gene pair correlation.

Complementary, there were five components in this cluster that only consisted of individual gene-gene pairs (Figure 28). The genes involved in these five components are listed in Table 4. Of these genes, there are two growth factor receptors, several kinases, a growth factor and a ras-related substrate. Growth factors and kinases are essential for proper MAPK signaling, and their inclusion in this cluster shows their

importance. However, since these genes only pair off with one other gene, and not the rest of the genes in the cluster, they are not as likely to be drivers of tumorigenesis of colorectal cancer. Rather, it is more likely that these genes have similar expression profiles to other more relevant gene-gene pairs, and were therefore grouped in this cluster based on correlation profile alone and not biological importance.

Table 4. Table of genes involved in the five subcomponents from Figure 28.

Gene Symbol	Full Gene Name
FGFR3	fibroblast growth factor receptor 3
FGFR4	fibroblast growth factor receptor 4
MAP3K11	mitogen-activated protein kinase kinase kinase 11
MKNK2	MAP kinase interacting serine/threonine kinase 2
CACNB2	calcium channel, voltage-dependent, beta 2 subunit
RPS6KA5	ribosomal protein S6 kinase, 90kDa, polypeptide 5
PLA2G12A	phospholipase A2, group XIIA
EGF	epidermal growth factor
TAOK3	TAO kinase 3
RAC1	ras-related C3 botulinum toxin substrate 1

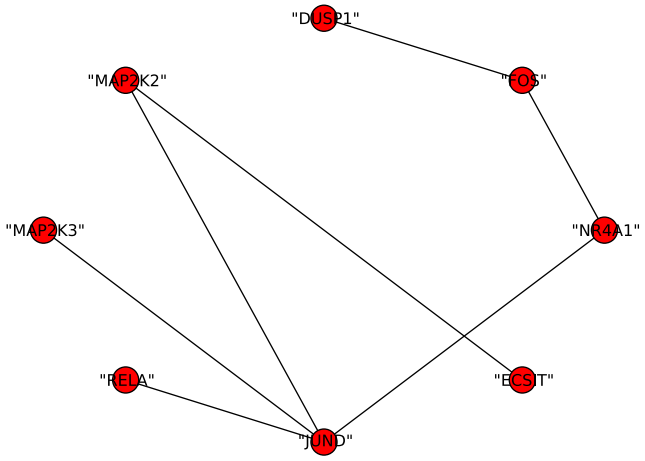


Figure 29. Network drawing of one component from partition.

Each node represents a gene and each edge represents a gene-gene pair correlation.

The last partition, seen in Figure 29, shows a group of eight genes that are connected with each other, and not connected to the other genes in the cluster, based on gene-gene pair correlations. The list of these genes is present in Table 5. From these genes, we can see there are several protein kinases and onco- and proto-oncogenes. The MAP kinases are essential for the signaling of the MAPK pathway, and interestingly are not connected to eachother, rather transiently connected via JUND, a proto-oncogene. This transient connetion between the two MAP kinases shows that they are more consistently regulated with JUND than each other, potentially showing that the regulation of these kinases is more dependent on the expression of JUND. Also, by having three oncogenes in this sub-component, there is added confidence that this set of genes is likely to be indicative for the tumorigenesis of colorectal cancer. As such, we propose that this set of eight genes

could potentially be used as a network biomarker for staging and predicting colorectal cancer. The correlation of these genes is distinct and separate from all other genes in the MAPK pathway, and by reducing the sample space to these eight genes, future experiments can be prioritized, and as a consequence performed more efficiently.

Table 5. Table of genes involved in the subcomponent from Figure 29.

Gene Symbol	Full Gene Name
RELA	v-rel reticuloendotheliosis viral oncogene homolog A (avian)
JUND	jun D proto-oncogene
MAP2K3	mitogen-activated protein kinase kinase 3
MAP2K2	mitogen-activated protein kinase kinase 2
ECSIT	ECSIT homolog (Drosophila) (evolutionarily conserved signaling intermediate in Toll pathways)
NR4A1	nuclear receptor subfamily 4, group A, member 1
FOS	FBJ murine osteosarcoma viral oncogene homolog
DUSP1	dual specificity phosphatase 1

5.4 Conclusions

During oncogenesis and tumorigenesis, many abnormalities arise within a cell, and the number of combinations of mutations and changes could potentially be innumerable. Such anomalies could consist of changes in gene expression, SNPs, protein mutations, fusion proteins, etc. Therefore, the identification of a specific aberration to associate to a disease stage and progression is no longer trivial. Alternatively, we look at a pathway level analysis, and by using the pathway correlation profiles, we can get a more general view of where in the pathways the dysregulation is occurring, regardless of the cause of

that dysregulation. From this approach, we can prioritize sets of genes as being responsible for the aberration of the pathway which will help prioritize future experiments for validation.

Here we looked into the Ribosome and MAPK Pathways with respect to the progression of colorectal cancer. We have been able to systematically isolate several oncogenes and tumor suppressor genes that have previously been implicated with colorectal and other cancers, both in diagnosis and metastasis. Also, we were able to make some novel predictions towards potential network biomarkers for identifying the stages of colon cancer. By not forcing genes into one cluster, like many previous methods do, we were able to computationally predict sets of genes that are responsible and involved in the pregression of this cancer, and these sets of genes better align with the underlying biology of these pathways. From this analysis, we feel we have shown the advantages to considering an alternative view of the data by using the pathway correlation profiles as an input for clustering, instead of the standard gene expression profiles that are frequently used.

As a result, using our method allows us to identify and prioritize gene targets of interest as related to colorectal cancer. We do note that not all our results were consistent with previously published studies. For example, finding decreased expression of SRF at metastasis. One potential reason for this is that we analyze the gene expression data, which does not always correlate with the functional data within a cell (i.e. metabolite levels, functional mutations, protein levels). Due to this limitation of our approach presented here, further experiments would need to be performed for validation of the accuracy of a predicted gene as well as identification of the reason for its involvement.

The approach presented here presents genes and gene sets that are likely to be involved in the disease progression (or experimental conditions). Regardless of the predictions and prioritizations derived from using the pathway correlation profiles as an input for clustering gene-gene pairs, further analysis and interpretation is essential for understanding why these genes were distinguished, and others were not.

6 Chapter 6: Discussion and conclusion

6.1 Summary

In this dissertation, I have presented the novel pathway correlation profile method for identifying and ranking pathways based on their perturbations. Many previous methods utilize a single gene approach, as well as an aggregate statistic for each gene. These previous methods fail to utilize all the biological data available, i.e., replicates, and potential added information gained from the interplay between genes, thereby taking a conservative view on pathway analysis. Our newly developed method looks into this orchestration between two genes across an entire pathway, does not depend on differential expression of any individual gene, utilizes biological replicates and is robust to the inherent noise of microarrays. Through this method, I can accumulate small perturbations in regulation over an entire pathway, subsequently ranking the pathways based on the significance of their perturbations. This method can make more biologically relevant predictions for pathway rankings that do not depend on a significant differential expression level, as well as accounts for pathway size, thus reducing the biases towards larger sized pathways.

Unlike many other gene set enrichment methods, the advantages to the pathway correlation profile method do not stop at ranking and identifying significantly perturbed pathways. Using the calculated correlation profile for each pathway, we can now identify specific gene-gene pairs that are responsible for the perturbation of the pathway, thereby extrapolating genes and pairs of genes that are likely to be driving the perturbation within a pathway. Through this method, we can better understand on a level of pathway

dynamics what is occurring in a pathway for a given condition and differential gene-gene pair correlations over multiple conditions or tissues. Moreover, we have extended the application of this method to consider temporal and disease progression data. When using this approach, we were able to look at the changes in gene-gene pair correlations through progression of a disease, colorectal cancer. Taken all together, the pathway correlation profile method not only prioritizes pathways based on the significance of perturbation, but also allows for the identification of causal reasons for the perturbation.

6.2 Advantages to analysis using pathway correlation profiles

Like other pathway analysis methods, the pathway correlation profile approach also ranks pathways based on their significance. However, our method has several advantages compared to the previously existing methods. This method utilizes the biological replications that are present in so many datasets.

Many existing methods frequently use just the sufficient statistics from the data (i.e., mean and variance), and calculate the significance of a gene based on these values. By only looking at the summary statistics for each gene, you are losing all the added information you gain from the biological replications.

6.3 Method limitations

Our pathway correlation profile method also has some limitations. Like other computational approaches based on gene expression analysis only, our method does not include regulatory mechanisms that may not be reflected in gene expression data, such as protein translational control, post-translational modifications and kinetic control of biochemical reactions. These are real limitations of our approach, and as a result we

could potentially make false negative predictions of false positive predictions. For example, if a protein needs to be phosphorylated for proper function, it is not relevant to consider the gene expression since even high expression levels of this gene would not correspond to activity if the phosphorylation were not occurring. Conversely, if a protein was not properly being degraded, then low levels of expression does not correspond to low protein levels as well. These are just two examples in which this approach could make false predictions.

These issues of false predictions may be addressed by incorporating other types of data in the analysis. By including protein expression or metabolomic data, we would have a better understanding of the active molecules in the pathway. Though obtaining this data is difficult and expensive, it would allow for a stronger analysis of the pathway and result in more accurate predictions. In addition to addition additional data sources, we also plan to develop a general software tool or plugin for users to apply our method easily.

7 Chapter 7: Future directions

7.1 Incorporate multiple data sources

In the perspective of systems biology, considering more than one type of information about the cell is essential [91]. Many methods have integrated gene expression data with protein-protein interaction (PPI) networks, e.g., to identify functional modules [92,93,94]. Using one data type in lieu of another, for example gene expression levels as a proxy for protein expression, has been commonplace [95,96]. Methods for using multiple data sources in tandem are still not very advanced, consequently there are still more room for improvement from computational methods and predictions.

Combining sequence level, gene expression, protein expression, metabolomic and epigenomic data to make more comprehensive predictions will allow for a better understanding of the biological context. A possible way to accomplish this systems approach could possibly be to create pathway correlation profiles for each data type individually, and then merging the results in the end. There are deficiencies with this approach, however, in that the results will miss interactions across data sets. For example, metabolites and proteins work in tandem to perform signaling. Alternatively, epigenomic modifications, such as DNA methylation, impede gene expression. Thus data sets need to be considered in tandem. To overcome these dependencies between the data types, calculating a pathway correlation profile that takes into account all pair-wise correlations (gene-gene, gene-protein, protein-epigenomic, etc.) could identify all true relations. A general limitation, however, is a global lack of data being available for a comprehensive analysis such as this.

7.2 Explore temporal changes

Pathway and network type analyses frequently look at single state models, or a difference between two states. To better understand the dynamics of pathways and signaling within a cell, it is essential to identify the perturbations in regulation for different tissue types, under varying treatments and conditions, and especially over time [97,98,99]. In many instances, complex diseases are not acute; rather, they are progressive. Identifying these temporal changes through the disease progression by developing computational and statistical methods to model temporal pathway dynamics will allow for a more complete understanding of the disease progression. In this dissertation, we looked at changes in select pathways in the progression of colorectal cancer, but more analytical and methodical approaches need to be developed to fully optimize the potential available from using pathway correlation profiles. Also, these methods can be used to better understand aging, embryogenesis, development, growth and many other aspects of physiology and pathology. This approach offers many opportunities for alternate therapeutics and treatments for diseases and many other avenues of implementation.

7.3 Classify genes based on regulation type

Currently, the pathway correlation profile method presented here utilizes all gene-gene pair combinations within a pathway. We ignore the structure of the pathway because that limits the potential for novel discoveries present in the data. In addition, we do not classify gene-gene pairs into groups, such as highly correlated, not correlated or negatively correlated, all of which carry their own biological significance. For example, if two genes have a high correlation, this suggests that there is a common mechanism of regulation between these two genes. This mechanism could be direct (e.g., gene A

activates gene B), indirect (e.g., gene A and gene B could be controlled by the same transcription factor), or another unidentified mean of regulation. Also, if two genes have no correlation, then they are assumed to be a background pair of genes (See Figure 30 for an example).

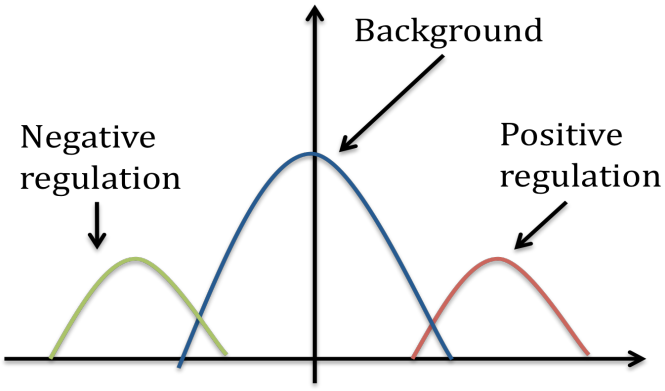


Figure 30. Example of a correlation profile classified by potential modes of gene regulation.

As shown in Figure 30, we can expect a correlation profile to be a mixture model of up to three normal distributions. Through this, we could potentially use statistical methods to model the correlation profile accordingly. After a model is developed, we can classify gene-gene pairs into their respective mode of regulation. Once we classify the gene-gene pairs, we can ask several questions: How many gene-gene pairs change mode of regulation? Is there a selective set of genes that maintain one mode of regulation? Do our predictions change when we remove background or uncorrelated gene-gene pairs? Through this approach, we can target different and more specific biological questions about the mode of regulation within a pathway, than we could through previous methods.

7.4 Pathway rewiring and driver mutations

Many cancers are thought to be driven by select mutations and potentially resulting in pathway rewiring to induce oncogenic signaling [77,100]. Alternatively, cancers undergo many mutations, some of which are necessary and implicated in cancer development (driver mutations) and some somatic mutations that have no known contribution to oncogenesis [101]. These abnormalities are hard to deduce from all the noise available in cancer data. Using this approach, we can help to tease out potential targets that could be driver mutations associated with oncogeneisis through significant changes in the correlation profile via a specific or few target genes. Additionally, mapping locations of perturbation back onto the pathway map could potentially spot loci of pathway rewiring. Both of these predictions can then be further analyzed in the wet lab. Being able to prioritize targets responsible for a phenotype will allow for more efficient, effective and systematic validation through wet lab experiments.

7.5 Network biomarkers for disease diagnosis

Establishing biomarkers for a disease is not a new concept [102]. Many methods have focused on identifying mutations [103], altered gene expression [104], microRNA expression [105], and aberrant DNA methylation [106] as biomarkers for a specific disease. Oftentimes, not solely one gene can identify a disease; to overcome this problem, studies have moved towards using several genes in combination as a network biomarker for diseases [107].

Using the pathway correlation profiles, we have the potential to identify network biomarkers. To accomplish this, we can select a set of significant genes that are most

responsible, and temporally discriminating, for the perturbation of a pathway. The combination of these genes, and their coordinated behavior, could not only identify a disease, but also determine the stage of the disease. These biomarkers can be robust to staging of diseases, and also try to model the pivotal biochemical changes through the progression of a disease.

7.6 Develop tool and plug-in

Creating new computational methods is necessary for advancing predictions. Developing a tool that will allow others to use your method is essential for general adaptation of the pathway correlation profile approach. For this, an R-package [108] or Cytoscape plug-in [109] can be developed and available for the general use. Through these tools, a naïve gene expression analysis would be performed and a pathway correlation profile analysis would be completed, resulting in a ranking of significant pathways. The database of pathways could be selected from an available list, such as KEGG [39], Reactome [48], TRANSPATH [49], Gene Ontology [110], etc., or the user could predefine the pathway database. Moreover, additional features in this tool could include additional analysis within a pathway, such as the temporal clustering presented here. Having a tool available would allow for this novel gene-gene pair correlation approach to really be accepted and utilized to its full potential.

8 References

1. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4: 44-57.
2. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102: 15545-15550.
3. Kim S-Y, Volsky D (2005) PAGE: Parametric Analysis of Gene Set Enrichment. BMC bioinformatics 6: 144.
4. Ackermann M, Strimmer K (2009) A general modular framework for gene set enrichment analysis. BMC bioinformatics 10: 47.
5. Song S, Black M (2008) Microarray-based gene set analysis: a comparison of current methods. BMC bioinformatics 9: 502.
6. Mootha V, Lindgren C, Eriksson K, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267 - 273.
7. Keller A, Backes C, Gerasch A, Kaufmann M, Kohlbacher O, et al. (2009) A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. Bioinformatics 25: 2787-2794.
8. Staal FJ, van der Burg M, Wessels LF, Barendregt BH, Baert MR, et al. (2003) DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 17: 1324-1332.
9. Yang JJ, Bhojwani D, Yang W, Cai X, Stocco G, et al. (2008) Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia. Blood 112: 4178-4183.
10. Adewale AJ, Dinu I, Potter JD, Liu Q, Yasui Y (2008) Pathway analysis of microarray data via regression. Journal of computational biology : a journal of computational molecular cell biology 15: 269-277.
11. Li K-C (2002) Genome-wide coexpression dynamics: Theory and application. Proceedings of the National Academy of Sciences 99: 16875-16880.
12. Horvath S, Dong J (2008) Geometric Interpretation of Gene Coexpression Network Analysis. PLoS Comput Biol 4: e1000117.
13. D,Äôhaeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16: 707-726.
14. Childs KL, Davidson RM, Buell CR (2011) Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PloS one 6: e22196.
15. Novak BA, Jain AN (2006) Pathway recognition and augmentation by computational analysis of microarray expression data. Bioinformatics 22: 233-241.
16. Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC bioinformatics 5: 18.

17. Lai Y, Wu B, Chen L, Zhao H (2004) A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics 20: 3146-3155.
18. Cho SB, Kim J, Kim JH (2009) Identifying set-wise differential co-expression in gene expression microarray data. BMC bioinformatics 10: 109.
19. Freudenberg JM, Sivaganesan S, Wagner M, Medvedovic M (2010) A semi-parametric Bayesian model for unsupervised differential co-expression analysis. BMC bioinformatics 11: 234.
20. D'haeseleer P (2005) How does gene expression clustering work? Nature biotechnology 23: 1499-1502.
21. Jain AK, Dubes RC (1988) Algorithms for clustering data: Prentice-Hall, Inc.
22. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95: 14863-14868.
23. Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. Nature genetics 32 Suppl: 502-508.
24. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America 96: 2907-2912.
25. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 97: 262-267.
26. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics 18 Suppl 1: S136-144.
27. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22: 1122-1129.
28. Li Y, Agarwal P, Rajagopalan D (2008) A global pathway crosstalk network. Bioinformatics 24: 1442-1447.
29. KRAUSE R, WILD DL. Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model; 2006. pp. 231-242.
30. Ernst J, Bar-Joseph Z (2006) STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics 7: 191.
31. Ernst J, Nau GJ, Bar-Joseph Z (2005) Clustering short time series gene expression data. Bioinformatics 21 Suppl 1: i159-168.
32. Hafemeister C, Costa IG, Schonhuth A, Schliep A (2011) Classifying short gene expression time-courses with Bayesian estimation of piecewise constant functions. Bioinformatics 27: 946-952.
33. Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19 Suppl 1: i255-263.
34. Bathoorn R, Welten M, Richardson M, Siebes A, Verbeek F (2010) Frequent episode mining to support pattern analysis in developmental biology. Pattern Recognition in Bioinformatics: 253-263.

35. Belmamoune M, Potikanond D, Verbeek FJ (2010) Mining and analysing spatio-temporal patterns of gene expression in an integrative database framework. J of Integrative Bioinformatics 7: 128.
36. Maurer LM, Yohannes E, Bondurant SS, Radmacher M, Slonczewski JL (2005) pH Regulates Genes for Flagellar Motility, Catabolism, and Oxidative Stress in Escherichia coli K-12. J Bacteriol 187: 304-319.
37. Singh J, Kumar D, Ramakrishnan N, Singhal V, Jervis J, et al. (2005) Transcriptional Response of Saccharomyces cerevisiae to Desiccation and Rehydration. Appl Environ Microbiol 71: 8752-8763.
38. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671-679.
39. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27-30.
40. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic acids research 31: e15.
41. Dennis G, Sherman B, Hosack D, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology 4: P3.
42. Streit WR, Entcheva P (2003) Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. Applied microbiology and biotechnology 61: 21-31.
43. Cronan JE, Jr. (1988) Expression of the biotin biosynthetic operon of Escherichia coli is regulated by the rate of protein biotination. The Journal of biological chemistry 263: 10332-10336.
44. Raser JM, O'Shea EK (2004) Control of Stochasticity in Eukaryotic Gene Expression. Science 304: 1811-1814.
45. Singh J, Kumar D, Ramakrishnan N, Singhal V, Jervis J, et al. (2005) Transcriptional response of Saccharomyces cerevisiae to desiccation and rehydration. Applied and environmental microbiology 71: 8752-8763.
46. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, et al. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 9: 166-180.
47. Sybesma W, Starrenburg M, Tijsseling L, Hoefnagel MH, Hugenholtz J (2003) Effects of cultivation conditions on folate production by lactic acid bacteria. Applied and environmental microbiology 69: 4542-4548.
48. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic acids research 33: D428-D432.
49. Krull M, Pistor S, Voss N, Kel A, Reuter I, et al. TRANSPATH¬Æ: an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucleic acids research 34: D546-D551.
50. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics 10: 57-63.
51. Alberts B Molecular Biology of the Cell in Cell 4th. Figure.
52. Ramakrishnan V (2002) Ribosome structure and the mechanism of translation. Cell 108: 557-572.

53. Poehlsgaard J, Douthwaite S (2005) The bacterial ribosome as a target for antibiotics. Nature reviews Microbiology 3: 870-881.
54. Fracasso G, Stirpe F, Colombatti M (2010) Ribosome-Inactivating Protein-Containing Conjugates for Therapeutic Use. Toxic Plant Proteins: 225-263.
55. Shirasaki T, Honda M, Mizuno H, Shimakami T, Okada H, et al. (2010) La Protein Required for Internal Ribosome Entry Site,ÄîDirected Translation Is a Potential Therapeutic Target for Hepatitis C Virus Replication. Journal of Infectious Diseases 202: 75-85.
56. Rison SC, Thornton JM (2002) Pathway evolution, structurally speaking. Current opinion in structural biology 12: 374-382.
57. Mano A, Tuller T, Beja O, Pinter RY (2010) Comparative classification of species and the study of pathway evolution based on the alignment of metabolic pathways. BMC bioinformatics 11 Suppl 1: S38.
58. Clemente JC, Satou K, Valiente G (2007) Phylogenetic reconstruction from non-genomic data. Bioinformatics 23: e110-115.
59. Koffas MA (2009) Expanding the repertoire of biofuel alternatives through metabolic pathway evolution. Proceedings of the National Academy of Sciences of the United States of America 106: 965-966.
60. Mazurie A, Bonchev D, Schwikowski B, Buck GA (2008) Phylogenetic distances are encoded in networks of interacting pathways. Bioinformatics 24: 2579-2585.
61. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. Trends in biochemical sciences 27: 514-520.
62. Ramagopal S, Ennis HL (1981) Regulation of synthesis of cell-specific ribosomal proteins during differentiation of Dictyostelium discoideum. Proceedings of the National Academy of Sciences of the United States of America 78: 3083-3087.
63. Reischl J, Schwenke S, Beekman JM, Mrowietz U, Sturzebecher S, et al. (2007) Increased expression of Wnt5a in psoriatic plaques. The Journal of investigative dermatology 127: 163-169.
64. Graham K, de las Morenas A, Tripathi A, King C, Kavanah M, et al. (2010) Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. British journal of cancer 102: 1284-1293.
65. Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT, Webster MJ, et al. (2006) Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Molecular psychiatry 11: 965-978.
66. Gutierrez A, Jr., Tschumper RC, Wu X, Shanafelt TD, Eckel-Passow J, et al. (2010) LEF-1 is a prosurvival factor in chronic lymphocytic leukemia and is expressed in the preleukemic state of monoclonal B-cell lymphocytosis. Blood 116: 2975-2983.
67. Hendrix ND, Wu R, Kuick R, Schwartz DR, Fearon ER, et al. (2006) Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas. Cancer research 66: 1354-1362.

68. Lenburg ME, Liou LS, Gerry NP, Frampton GM, Cohen HT, et al. (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. BMC cancer 3: 31.
69. Kondrashov N, Pusic A, Stumpf CR, Shimizu K, Hsieh AC, et al. (2011) Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. Cell 145: 383-397.
70. Stoneley M, Willis AE (2004) Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. Oncogene 23: 3200-3207.
71. Knudson AG (2001) Two genetic hits (more or less) to cancer. Nature reviews Cancer 1: 157-162.
72. Nicholson JK (2006) Global systems biology, personalized medicine and molecular epidemiology. Molecular systems biology 2: 52.
73. Fang JY, Richardson BC (2005) The MAPK signalling pathways and colorectal cancer. The lancet oncology 6: 322-327.
74. Dunn KL, Espino PS, Drobic B, He S, Davie JR (2005) The Ras-MAPK signal transduction pathway, cancer and chromatin remodeling. Biochemistry and cell biology = Biochimie et biologie cellulaire 83: 1-14.
75. Sebolt-Leopold JS, Dudley DT, Herrera R, Van Becelaere K, Wiland A, et al. (1999) Blockade of the MAP kinase pathway suppresses growth of colon tumors in vivo. Nature medicine 5: 810-816.
76. Gulmann C, Sheehan KM, Conroy RM, Wulfkuhle JD, Espina V, et al. (2009) Quantitative cell signalling analysis reveals down-regulation of MAPK pathway activation in colorectal cancer. The Journal of pathology 218: 514-519.
77. Pawson T, Warner N (2007) Oncogenic re-wiring of cellular signaling pathways. Oncogene 26: 1268-1275.
78. Sheffer M, Bacolod MD, Zuk O, Giardina SF, Pincas H, et al. (2009) Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proceedings of the National Academy of Sciences of the United States of America 106: 7131-7136.
79. Compton CC, Greene FL (2004) The staging of colorectal cancer: 2004 and beyond. CA: a cancer journal for clinicians 54: 295-308.
80. Camps J, Nguyen QT, Padilla-Nash HM, Knutsen T, McNeil NE, et al. (2009) Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. Genes, chromosomes & cancer 48: 1002-1017.
81. Ruggero D, Pandolfi PP (2003) Does the ribosome translate cancer? Nature reviews Cancer 3: 179-192.
82. Hoeller D, Hecker CM, Dikic I (2006) Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. Nature reviews Cancer 6: 776-788.
83. Hoeller D, Dikic I (2009) Targeting the ubiquitin system in cancer therapy. Nature 458: 438-444.
84. Yang Y, Kitagaki J, Wang H, Hou DX, Perantoni AO (2009) Targeting the ubiquitin-proteasome system for cancer therapy. Cancer science 100: 24-28.
85. Kirkin V, Dikic I (2011) Ubiquitin networks in cancer. Current opinion in genetics & development 21: 21-28.

86. Mancini M, Toker A (2009) NFAT proteins: emerging roles in cancer progression. Nature reviews Cancer 9: 810-820.
87. Patil S, Chamberlain RS (2012) Neoplasms Associated with Germline and Somatic NF1 Gene Mutations. The Oncologist.
88. Fletcher O, Houlston RS (2010) Architecture of inherited susceptibility to common cancer. Nature reviews Cancer 10: 353-361.
89. Choi H, Kim K, Lee J, Park H, Jang K, et al. (2009) Serum response factor enhances liver metastasis of colorectal carcinoma via alteration of the E-cadherin/beta-catenin complex. Oncol Rep 21: 57-63.
90. M√ºller MR, Rao A (2010) NFAT, immunity and cancer: a transcription factor comes of age. Nature Reviews Immunology 10: 645-656.
91. Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nature reviews Molecular cell biology 7: 198-210.
92. Ewing RM, Chu P, Elisma F, Li H, Taylor P, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. Molecular systems biology 3: 89.
93. Lahti L, Knuuttila JE, Kaski S (2010) Global modeling of transcriptional responses in interaction networks. Bioinformatics 26: 2713-2720.
94. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nature biotechnology 23: 561-566.
95. Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. Genome biology 4: 117.
96. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature biotechnology 25: 117-124.
97. Schmid PR, Palmer NP, Kohane IS, Berger B (2012) Making sense out of massive data by going beyond differential expression. Proceedings of the National Academy of Sciences of the United States of America 109: 5594-5599.
98. Valcarcel B, Wurtz P, Seich al Basatena NK, Tukiainen T, Kangas AJ, et al. (2011) A differential network approach to exploring differences between biological states: an application to prediabetes. PloS one 6: e24702.
99. Ouyang Z, Song M, Guth R, Ha TJ, Larouche M, et al. (2011) Conserved and differential gene interactions in dynamical biological systems. Bioinformatics 27: 2851-2858.
100. Hsu PP, Sabatini DM (2008) Cancer cell metabolism: Warburg and beyond. Cell 134: 703-707.
101. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719-724.
102. Srinivas PR, Kramer BS, Srivastava S (2001) Trends in biomarker research for cancer detection. The lancet oncology 2: 698-704.
103. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, et al. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research 69: 6660-6667.
104. Sotiriou C, Pusztai L (2009) Gene-expression signatures in breast cancer. The New England journal of medicine 360: 790-800.

105. Chen X, Ba Y, Ma L, Cai X, Yin Y, et al. (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell research 18: 997-1006.
106. Kobayashi Y, Absher DM, Gulzar ZG, Young SR, McKenney JK, et al. (2011) DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. Genome research 21: 1017-1027.
107. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Molecular systems biology 3: 140.
108. Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. Journal of computational and graphical statistics 5: 299-314.
109. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13: 2498-2504.
110. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25: 25.

Vita

Allison Tegge was born in Illinois. After high school, she attended the University of Illinois Urbana-Champaign where she received her B.S. in Animal Sciences. Upon completion in 2006, she continued and earned her M.S. in Bioinformatics from the Animal Sciences department at the University of Illinois Urbana-Champaign. Allison continued on in her studies at the University of Missouri.

Allison Tegge's research interests include computational and systems biology, pathway analysis and cancer informatics. She received her PhD in Informatics from the University of Missouri Informatics Institute in December 2012.