Evaluating the Consistency of Gene Set Sources for Use in Pathway Analysis

Alexandra Sitarik
Department of Mathematics, Wittenberg University
April 12, 2011

Abstract

Pathway Analysis, or gene set analysis, is a fairly new and promising approach to Genome-Wide Association Studies. Not only does this method avoid the multiple testing penalties associated with SNP-by-SNP analyses, but it also incorporates *a-priori* biological knowledge about the interactions of genes in pathways. There are many sources that produce these sets for implementation in gene set analysis. However, these sources all create their sets using different methods, based on varying principles and using different underlying algorithms. The importance of the consistency of these gene sets cannot be disregarded, as inconsistent sets produce inconsistent genotypic/phenotypic associations in gene set analysis. The goal of this study was to compare a number of these commonly used sources of sets to see which produces the most consistent sets of genes. Gene expression microarray data was used for this evaluation.

Background

Genome-Wide Association Studies (GWAS) seeks to identify the genetic components of complex diseases.[1-4] Since the explosion of popularity in such studies in the last decade, many different methods have been proposed to accomplish these endeavors efficiently and accurately. Perhaps one of the most promising methods at our disposal thus far is the method of Pathway Analysis.[1(p1)] Unlike many other methods that only look for genotype-phenotype association at the single SNP level[*] or at the single gene level, Pathway Analysis looks for association across biologically related, interconnected sets. These sets begin by associating SNPs to genes, and then by creating related sets

[*] A Single Nucleotide Polymorphism, or SNP, is a single variation in the DNA sequence at a single nucleotide.

of genes. Each of these sets is summarized based on their association, and the statistical significance of that association is assessed at each step in the process.[5]

One of the most beneficial aspects of Pathway Analysis is that each of the steps mentioned above (SNP aggregation, gene aggregation, etc) can be individually analyzed and optimized, thereby contributing to the accuracy of the method as a whole. This valuable feature provides the motivation for this project: to zoom in on one step of the analysis and seek to improve it. Specifically, this research focuses on the gene aggregation step, or how genes are grouped together into sets of genes and how this affects the overall analysis.

On an individual gene level, it is difficult to detect a significant connection between the gene and a given phenotype, due to the vast size of a genome weakening this connection. Since most mathematical models assume genes show a similar, but weak connection with the phenotype, we use methods to aggregate genes, which then create a statistically significant connection with the phenotype because the connection is strengthened. If we combine genes in a significant way using previously determined biological knowledge, the assumption about the individual connections of genes with the phenotype is met. Our basic goal then, is to test whether we are in fact aggregating genes in a significant way. To achieve this, we measured the correlation in phenotype patterns between genes in the same gene set to see if they had consistent effects on their expression patterns.

Ideally, we wish for genes in a gene set to be “consistent”, or for all genes in a gene set to be functionally related to one another. If genes are functionally related to one another, we will obtain accurate associations with the phenotype of interest in Pathway Analysis. As an example, suppose I group gene A and gene B together in a gene set and suppose I find that gene A has association with the phenotype. Pathway Analysis then, assumes that gene B also has some kind of association with the phenotype, because gene A and gene B are functionally related. If they are in fact not functionally related, this process falls apart. This “consistency” between genes in a gene set is exactly what we need to test for.

Pathway Analysis is a complicated process with much room for variation and error in each of the steps. Though our analysis is only focused on “the gene set step”, there are still many factors that could affect the accuracy of phenotype association at the gene set level in the human genome. Because of this complication, it became necessary to find a way to simply focus on the methods by which gene sets are created for each source, while still controlling for the other factors and more complicated interactions that come into play in the human genome. For this reason, we turned to microbial organisms.

Because microbial genes have fewer SNPs and there are fewer genes overall, they provide us with an excellent comparison to human gene sets in our analysis. Surprisingly--and to our advantage--there are many parallels that can be drawn between microbial gene sets and human gene sets that can help us to understand human gene sets. First note that the sources of sets we are analyzing use comparable methods for both human and microbial gene sets (with the exception of Operons as they do not occur in human genes), which is the primary groundwork by which we can make this modification. Further, gene expression data is obtained in a similar manner on microbial and human genes. This concept is key, as gene expression data acts as our main tool of set comparison (gene expression data will be discussed in detail in the Materials section). If gene expression data differed, we would have no basis to draw parallels between human and microbial sets. In addition, Pathway Analysis can be applied to microbial organisms easily and successfully, making the effects of the gene sets on the overall Pathway Analysis a relevant point to discuss, both for those interested in the human genome and those interested in the microbial genome.

Materials

Set Sources

When conducting a Pathway Analysis, previously created gene sets that incorporate *a priori* biological information are often used in the process. There are many different set sources that are frequently used that produce these gene sets. All of these set sources aggregate gene to create gene sets using various concepts about the interaction of genes in a genome. We chose to evaluate six of these set sources in order to determine which of them creates the most consistent gene sets. The set sources we selected are: The Gene Ontology,[6] KEGG (Kyoto Encyclopedia of Genes and Genomes),[7] Predicted Operons from Microbes Online,[8] Scenarios from The SEED, Scenario Paths from The SEED, and Subsystems from The SEED.[9] A description of the principles and processes by which these set sources form their gene sets are given below:

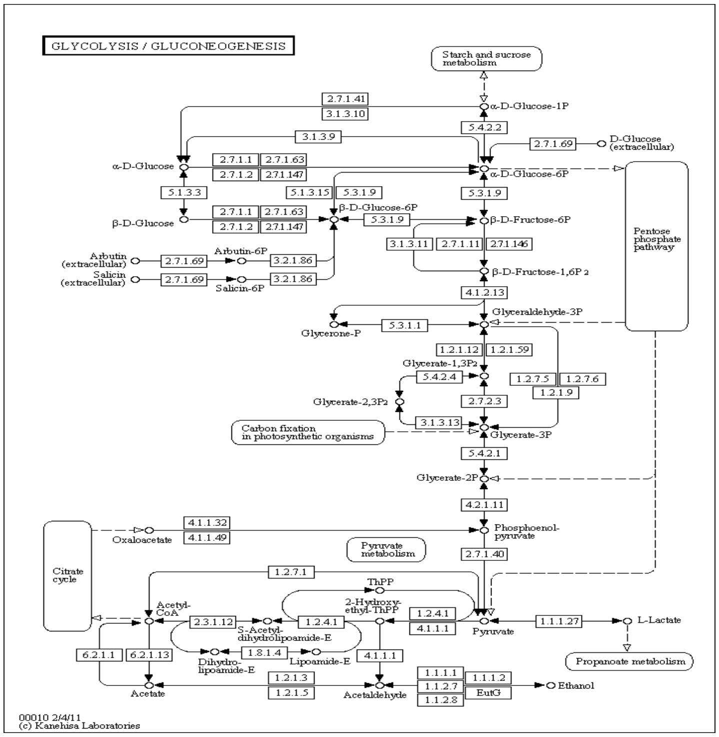
1) **The Gene Ontology (GO)** - forms gene sets based on similar biological processes, molecular functions, and cellular components. Comparisons are evaluated independently within each organism.[6]

2) **Subsystems**- use similar methods as Gene Ontology (groups genes by similar biological processes, molecular functions, cellular components, etcetera), but also groups genes based on similar gene products. Additionally, subsystems use comparative genomics for comparable organisms in order to capitalize on cross-organismal similarities.[9]

3) **Scenarios**- subsets of subsystems in which genes form a chain of connected reactions. Scenarios are extremely specified sets not only for these reasons, but also because they are strictly metabolic reaction networks that perform no other functions other than those involving metabolism.[9]

4) **Kyoto Encyclopedia of Genes and Genomes (KEGG)** - collections of genes intertwined by higher-level systematic functions, such as molecular interaction and reaction networks. All genes must be "connected" within a pathway, not just functionally related. KEGG Pathways are comparable to a scenario, but are much larger sets. KEGG represents their gene sets by using KEGG Pathway maps, such as the one in Figure 1[7]:

Figure 1



5) **Scenario Paths (Paths)** - subsets of scenarios (even more fine-tuned subsystems, scenarios). All genes in a path must be connected in terms of their gene products, and these connections are only considered a path if they are within a metabolic reaction network. Paths are comparable to a “mini-KEGG pathway”. That is, looking at the KEGG map above, a Path would be a specific set of connected genes within a KEGG map. [9]

6) **Predicted Operons**- by definition, an operon is a cluster of genes that must all be contiguous and under the control of the same single regulatory promoter, which essentially regulates the transcription of a given gene. Operons are unique to microbial organisms and are not found in most multi-cellular organisms. Microbes Online attempts to predict all operons for a given genome by taking into account other correlated factors along with contiguity, including: nucleotide distance between two genes and other functional categories defined by Microbes Online.[8]

Organisms

Gene sets were collected for every set source across six different microbial organisms. The organism names along with the abbreviations used here are as follows:

Staphylococcus aureus (Staph aureus), ***Escherichia coli*** (E. coli), ***Shewanella oneidensis*** (Shewanella), ***Thermus Thermophilus*** (Thermus), ***Bacteroides thetaiotaomicron*** (Bacteroides), and ***Pseudomonas aeruginosa*** (Pseudomonas). These particular organisms were selected for their biological diversity to allow for more general conclusions to be drawn about the consistency of set sources. A table is given below listing the number of total gene sets collected for each set source and each organism:

	GO	KEGG	Operons	Paths	Scenarios	Subsystems	Total
Staph aureus	975	90	522	267	139	360	2353
Pseudomonas	1170	104	1049	323	191	389	3226
Shewanella	1034	95	704	259	150	327	2569
Bacteroides	967	79	1009	235	137	193	2620
Thermus	853	82	419	251	138	178	1921
E. coli	1170	98	753	314	192	399	2926
Total	6169	548	4456	1649	947	1846	15615

Table 1: Total Number of Sets for Source and Organism

The important observation to make from the above table is that some set sources identify many more sets than other set sources, but a fixed set source generally identifies about the same number of sets across all organisms.

Gene Expression Microarray Data

Since the ultimate goal of this study is to see how consistent a gene set is, or how similar the genes are within a gene set, a tool was needed to provide measures of similarity between groups of genes. The tool we chose to analyze these similarities is gene expression microarray data. Essentially, a gene expression microarray is the result of a controlled experiment performed across the genome. This controlled experiment analyzes how a given gene reacts to that particular experiment, where this reaction is given by a measure of RNA produced per gene. RNA production is used as a measurement of “gene activity” and is used in many contexts to investigate genotype-phenotype association. The entire dataset, which consists of thousands of microarrays, was produced using Affymetrix GeneChips for each organism, which was generously provided by the Many Microbe Microarrays Database (M^{3D})[10] and the Gene Expression Omnibus.[11] The Affymetrix GeneChips provide raw CEL files, which were then background-corrected, normalized, and summarized using Robust Multichip Averaging. The dataset consists of a variety of arrays that seek to target a variety of experimental conditions and strains for each organism, in order to better understand the functions and processes associated with a given gene. Below is a table giving the number of genes that were analyzed in the microarray data per organism, and the total number of microarrays collected and analyzed for that organism:

Organism	Number of Genes	Number of Microarrays
Staph aureus	2885	852
Pseudomonas	5666	176
Shewanella	4529	245
Bacteroides	4913	41
Thermus	2260	407
E. coli	4468	907

Table 2: Number of Genes and Microarrays per Organism

Methods

The gene expression microarray data provides us with our necessary per gene measure of activity in order to quantify gene set consistency. Because the expression data is normalized as mentioned above, these measures are the $\log_2$ of the raw expression data, and typically range from 4 to16. For two particular fixed genes in a gene set of a given organism, we can now look across these hundreds of microarrays for that organism and ask: did these two particular genes react similarly to these controlled experiments? If the answer is yes, that they do correspond, (ie, they both produce high amounts of RNA for certain experiments and low amounts for others), then they are related to a similar group of functions that these controlled experiments coincide with. The next step is to find a way to statistically determine how similar their reactions really were. For this purpose, we calculated a measure of correlation with Pearson's pair-wise r correlation, where r is given by[12]:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

In this formula, X is the expression values for the first gene, Y is the expression values for the second gene, and n is the total number of expression values measured for both genes. Since the r correlation alone produces a value between -1 and 1, we chose to square this r to produce a value from 0 to 1. We did this because the negative or positive sign of r tells you about the direction of the correlation, but we are only interested in the magnitude, or how much they are correlated, and not the direction. This r^2 is the main value we use to establish correlation. We then find all possible pairs of genes in a gene set, which will be $\binom{m}{2}$ when there are m genes in a gene set, and similarly calculate all of their pair-wise r^2 values, resulting in $\binom{m}{2}$ r^2 values for that gene set. From this, we can draw basic statistics for a given gene set to express the overall correlation of the gene set as a whole, such as mean and median. These calculations were performed on all gene sets of a given set source, as long as the set had at least two members. These statistics were calculated using custom R scripts[13] and with the computational help of the Parallel Computing Cluster at Hope College.[14]

Statistical Analysis

I. Set Sizes

Before we discuss the analysis of the correlation values as calculated above, it is necessary to understand more basic and general trends in the data. Below is a table of average set sizes and their corresponding standard deviations, for each set source/organism combination:

	Set Source											
	GO		KEGG		Operons		Paths		Scenarios		Subsystems	
Organism	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Staph aureus	40.90	120.32	14.87	15.24	3.063	2.079	5.667	3.298	4.475	2.798	7.135	5.996
Pseudomonas	71.78	239.71	24.70	27.89	3.186	2.030	6.096	4.305	5.429	4.339	10.133	9.284
Shewanella	54.11	164.86	17.43	16.43	3.191	2.465	5.454	3.118	4.834	2.913	8.915	7.493
Bacteroides	58.72	184.85	15.32	11.53	3.401	2.167	6.489	4.025	4.876	3.431	8.000	8.867
Thermus	32.72	93.27	12.75	11.45	2.970	2.130	4.856	2.832	4.291	2.900	6.778	5.827
E. coli	56.10	185.54	20.69	24.42	3.200	2.183	5.975	4.143	4.823	3.147	8.681	7.486

Table 3: Set Sizes-Decomposed by Organisms and Set Source

Notice that some set sources tend to have larger sets on average, while some set sources have significantly smaller set sizes on average. In particular, notice Gene Ontology's (GO) typically extremely large sets and Operons' typically small sets. Roughly speaking, this tells us that set sources create gene sets of varying sizes in relation to one another. We can also see in this table by looking at a fixed set source across all organisms that gene sets tend to have approximately the same set sizes. This is reassuring for our analysis: it tells us that their methods are producing similarly sized gene sets, a hint that the methods are consistent across organisms. For extra statistical support, we can formally test to see if set source and organism have an effect on set sizes using Analysis of Variance (ANOVA).[15] Essentially, ANOVA is a statistical tool which compares differences in means across groups. Before such an analysis, however, there is one more important thing to notice from this table: high mean sets sizes correspond to high standard deviations (in particular, GO and KEGG). Before we should proceed using ANOVA, we should transform our data to better meet ANOVA test conditions. For this reason, for the remainder of our analysis we will be using the $\log_{10}$ of the set sizes. Here are the results of such an analysis:

Sources of variability	DF	Seq SS	Adj SS	Adj MS	F-statistic	p-value
Set Source	5	1064.938	1011.621	202.324	949.28	0.000
Org	5	18.444	11.306	2.261	10.61	0.000
Set Source*Org	25	7.456	7.456	0.298	1.40	0.089
Error	15680	3341.962	3341.962	0.213		
Total	15715	4432.800				

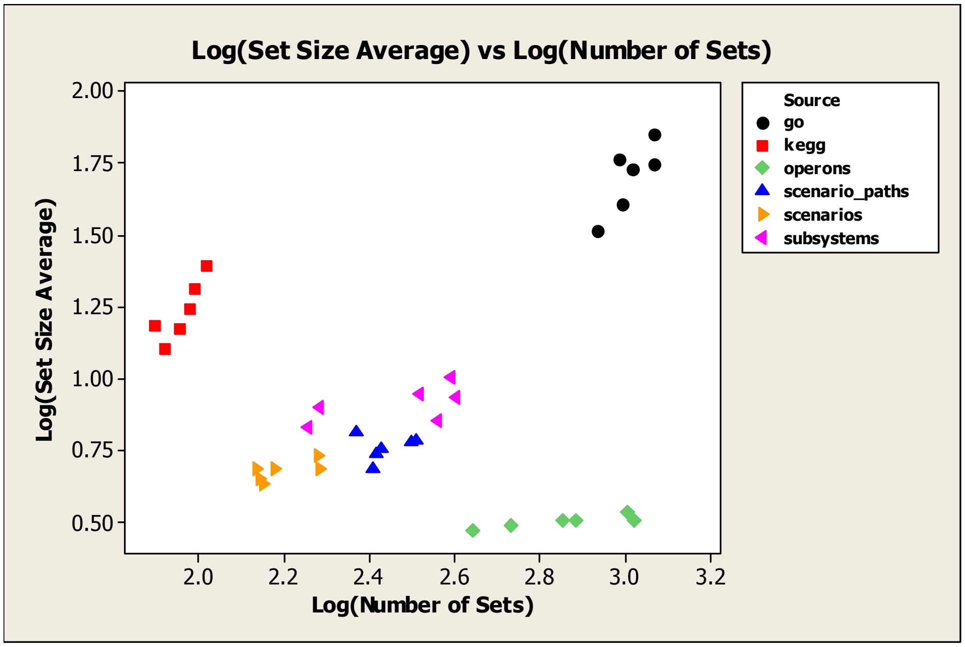
*Table 4: ANOVA Table for Log(Set Sizes), with Source, Organism, and Source*Organism as Factors of Interest*

S = 0.461666   R-Sq = 24.61%   R-Sq(adj) = 24.44%

Our ANOVA analysis confirms our intuitive observations. By the large F-statistic for set source, we now know that there is a lot of variability in set size from set source to set source. ANOVA also produced a significant F-statistic for organism, meaning there is also variability on set size due to organism. Notice set source and org are both supported by practically 0 p-values. Our confidence in set source*org interaction is much less. Notice the interaction factor not only has a fairly small F, but also a high p-value.

We can also look at the relationship between the total number of sets and the average number of genes per set for each set source. Here is a graph demonstrating this relationship (using the log of both set size average and number of sets):

Figure 2



To begin, some set sources produce a much larger number of sets for all organisms (ie, Gene Ontology). There are also some sets that produce a fewer number of sets (most clearly KEGG). Others are scattered between 200 and 400 sets per organism, with the exception of Operons which produce a wider variety of number of sets per organism. However, this is expected, as Operons are directly tied to the size of the genome. There is also another interesting trend here: the relationship of number of sets with set size average among sources. For Gene Ontology, this is a clear linear trend (as the number of sets increases, so does the size of each individual set). Others remain more constant. For example, even though Operons have the widest variety of the number of sets per organism, they are clearly the most constant in set size averages. This information is provided to simply highlight the many differences in how gene sets are formed across set sources to better understand the data in later analysis.

II. Correlation Values: Mean versus Median

We are now ready to discuss overall results of the correlation values, which were obtained as described in the Methods section. As previously mentioned, we calculated both mean and median correlation values for each set per organism, per set source. The question then arises, which of these is a better measure of the overall trend in the correlation values? Means can sometimes be misleading, as they are easily pulled down by small observations or pulled up by large observations. Figures 3 and 4 are histograms of both the mean and the median correlation values:

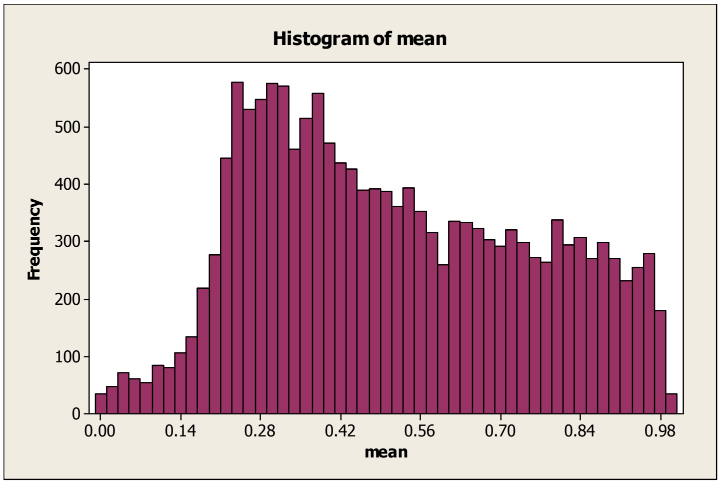


Figure 3

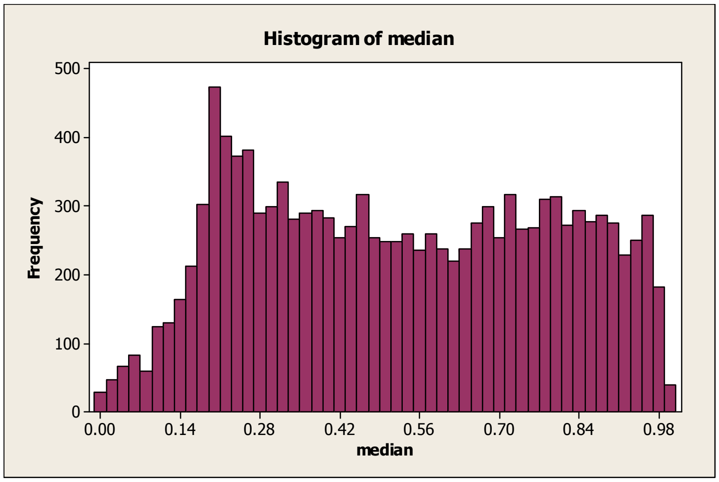


Figure 4

From these histograms, you can tell that the distributions are roughly the same, and it does not appear that the means are skewing the data in any misleading way. Even though these graphs tell us something about their distributions, they tell us nothing about how “correlated” the correlation values are. That is, do low mean correlation values correspond to low median correlation values and vice-versa? A scatter plot of mean versus median can help to answer this question:

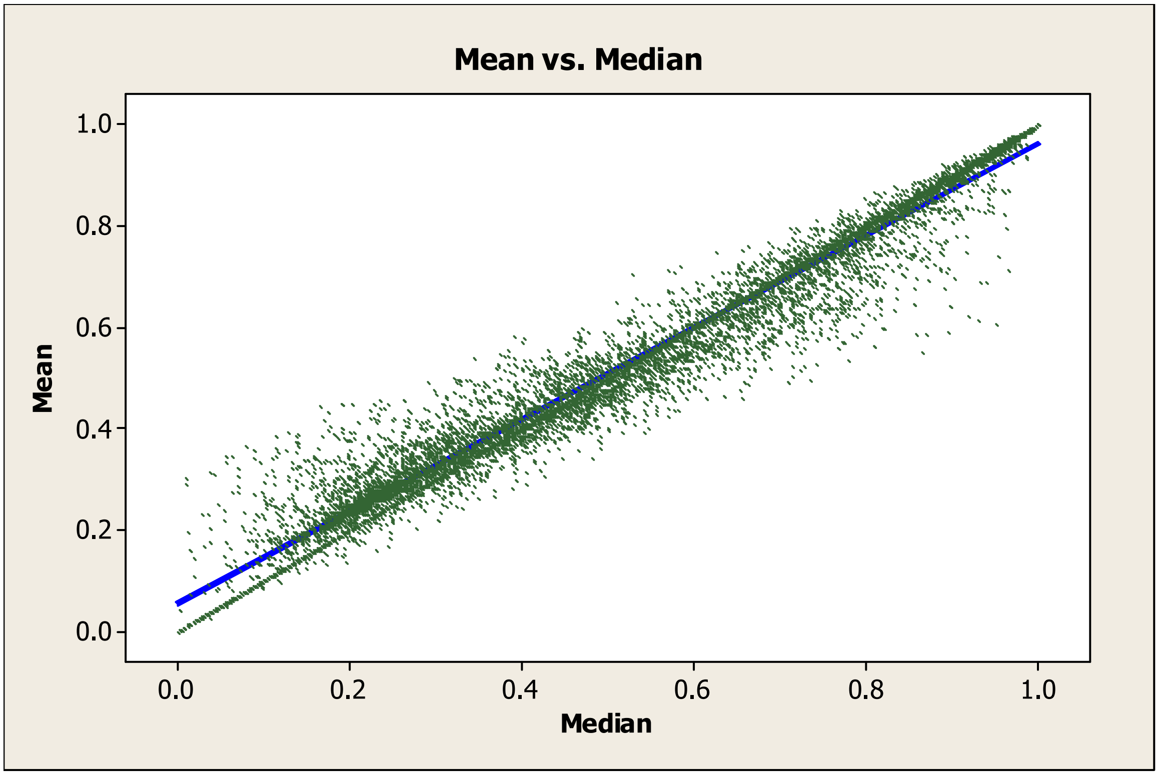


Figure 5

The regression line of this scatter plot is y = 0.908x + 0.0554, which just about has a slope of 1. This tells us that on average, median and mean r^2 values for a set produce about the same value, or that the median is no better of a measure than the median. Since we have established that in this analysis mean and median generate similar results, we will proceed using the mean only for the rest of this analysis, for simplicity’s sake.

III. Correlation by Set Source and Organism

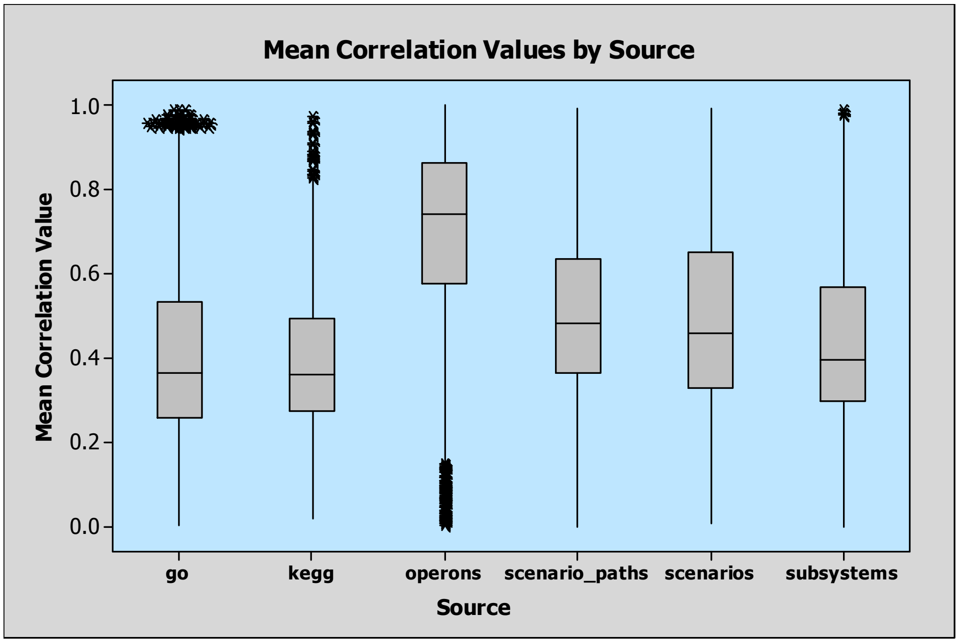
Now that we have an appropriate measure of our r^2 values, we can compare average values across all six set sources:

Source	Mean	SD
GO	0.41234	0.20310
KEGG	0.40941	0.18589
Operons	0.69758	0.21307
Paths	0.50257	0.21148
Scenarios	0.49437	0.22831
Subsystems	0.44452	0.20199

Table 5: Average Correlation Values by Set Source

From these very basic statistics, it appears as though Operons tend to have the highest r^2 values on average by a fairly significant amount. The next most correlated set source appears to be Paths, followed by Scenarios. The remainder of set sources has approximately equal average correlation values, centered near .45. Also notice the standard deviations for each set source, which are controlled and near .2. This is rather expected however, as the response correlation value is forced to be between 0 and 1. Here is a more descriptive visual of this information in the form of a box plot:

Figure 6



There are two important patterns to observe here. The first is the skewness of each set source. By the elongated right tails and larger right sides of the boxes, we see that nearly all set sources are skewed right, which means that most of the observations are concentrated on the low end of the correlation scale. In fact, Operons is the only set source for which this pattern does not hold, as they are in fact left-skewed. Even more interesting is the outliers for each of the set sources. Notice several Gene Ontology sets with high correlation values were recognized as outliers in the data, as did a few KEGG and Subsystem sets. This means that high correlation values were not expected to be in the data for these set sources. Nevertheless however, it should be noted that even though Gene Ontology has typically low correlation values, there are still several highly correlated sets. On the other hand, in the case of the Operons the *small* observations were recognized as outliers, which do not follow the expected pattern of the data. This is the only set source that has any outliers on the low end of the correlation values. Overall, this box plot further supports the consistency of Operons.

We can examine a similar visual of correlation values by organism:

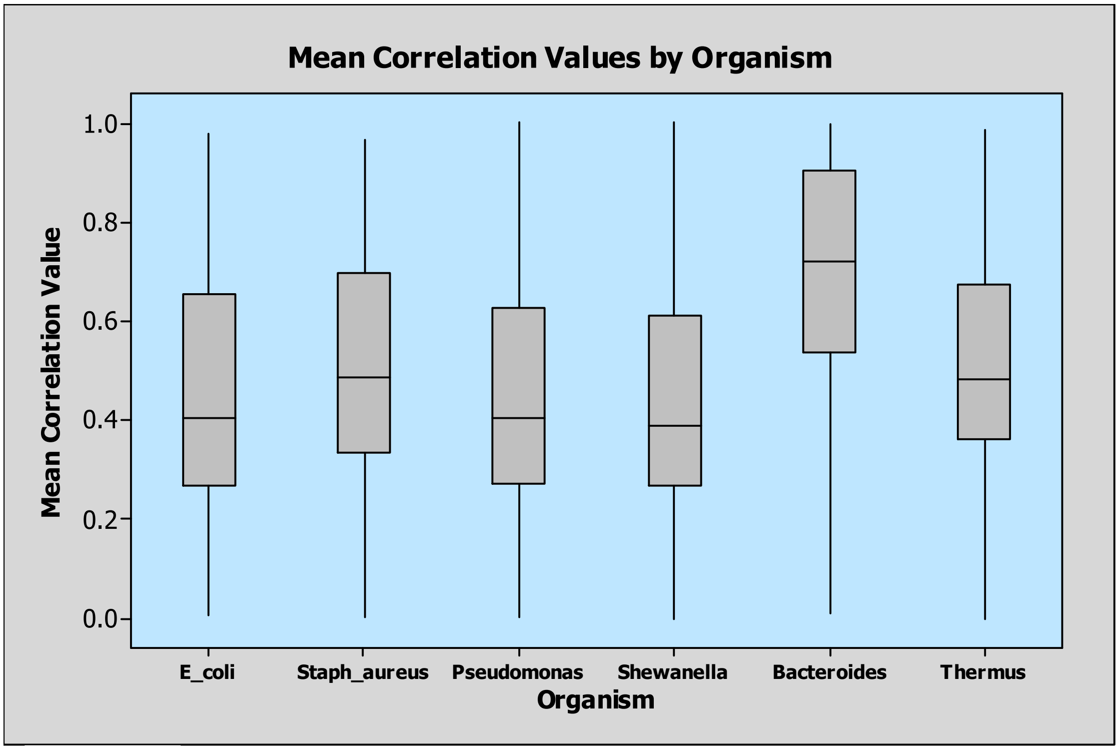


Figure 7

The data shows that there is much less variation on correlation by organism. Most organisms have similar typical values, have very similar spreads, and do not have outliers. In fact, Bacteroides seems to be the only organism expressing variation.

Besides correlation values broken down solely by set source or by organism, we can also examine how correlation values are affected by organism-to-organism differences, as well as differences by set source. It is important to not forget organism-to-organism variability in the design, as we are still unsure how much our methods differ across organisms. Below is a table showing correlation values decomposed by both set source and organism:

	GO		KEGG		Operons		Paths		Scenarios		Subsystems	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Staph aureus	0.408	0.165	0.380	0.153	0.724	0.176	0.568	0.186	0.537	0.212	0.4746	0.181
Pseudomonas	0.339	0.170	0.343	0.130	0.618	0.226	0.432	0.182	0.421	0.188	0.403	0.178
Shewanella	0.335	0.162	0.332	0.137	0.663	0.201	0.457	0.209	0.423	0.206	0.377	0.166
Bacteroides	0.634	0.193	0.682	0.172	0.786	0.218	0.659	0.204	0.703	0.218	0.671	0.225
Thermus	0.465	0.168	0.440	0.139	0.702	0.212	0.499	0.200	0.492	0.213	0.455	0.187
E. coli	0.336	0.181	0.336	0.138	0.702	0.171	0.444	0.201	0.445	0.216	0.399	0.182

Table 6: Correlation Values Decomposed by Organism and Set Source

There are a few interesting things to note from this table. The first is to notice that our previous observation about correlation values and set source still hold consistent when broken down by organism. That is, for a given organism, set sources still seemed to be “ranked” similarly, from best set source to poorer set sources. The second is to notice trends in organism-to-organism differences. For example, it appears that there tends to be higher correlation values for organism Bacteroides, as we noted in the box plot above. Reasons for this variation are less obvious. Most likely, this is due to the unusually low number of microarrays analyzed for this particular organism. Referencing Table 2, there were only 41 microarrays, while all other organisms had at least 175. There also could be some unexplained variation in the expression data for this organism. Regardless, the main point to observe is that set sources are not perfectly consistent across organisms, but how inconsistent they are is still unclear.

Now that we have formulated informal inferences about the effects of set source and organism on correlation value, we can formally test these inferences using ANOVA. In addition to set source, we will make organism a factor in the design to help to clarify the important matter of organism-to-organism variability discussed above. Below is the ANOVA table resulting from this analysis:

Sources of variability	DF	Seq SS	Adj SS	Adj MS	F-statistic	p-value
Set Source	5	229.35	200.83	40.17	1138.01	0.000
Org	5	106.58	61.18	12.24	346.67	0.000
Set Source*Org	25	17.45	17.45	.698	19.77	0.000
Error	15579	549.85	549.85	.035		
Total	15614	903.22				

*Table 7: ANOVA Table for Correlation Value, with Source, Org, and Source*Org as Factors of Interest*

S = 0.187868   R-Sq = 39.12%   R-Sq(adj) = 38.99%

Our ANOVA analysis confirms our suspicions about the effects of both set source and organism on correlation. By the large F-statistic for set source, we now know that there is about 1138 times more variability in correlation value from set source to set source than there is within an individual set source. Similarly, we also know that there is about 346 times more variability in correlation value from organism-to-organism than there is within an individual organism. There is also grounds to say there is set source*org interaction on correlation as well. The low p-values of our F-statistics further support all of these findings.

IV. Stratification by Set Size

When it comes to discovering which conditions have an effect on the correlation value of a gene set, another important question to ask is: what effect (if any) does set size have on correlation? A reasonable educated guess would be that sets that have fewer genes have a higher correlation value, because there are more interactions between genes in a large set, causing differences between them. To investigate this theory, we started with a simple table of all gene sets showing correlation by set size. Each set is grouped into a range, based on its corresponding set size, and the mean and standard deviation of those correlation values was calculated, as shown in Table 8:

Size Range	Mean	SD
2-5	0.56954	0.26045
6-10	0.48518	0.19852
11-20	0.43622	0.16908
21-40	0.39427	0.15703
41-60	0.36191	0.15451
61-80	0.3580	0.1451
81-120	0.34509	0.13400
121-160	0.3527	0.1305
161-250	0.3334	0.1295
251-350	0.3351	0.1415
351-450	0.3147	0.1174
451-550	0.2859	0.1031
551-650	0.2914	0.0978
651-1000	0.3224	0.1173
1001-2000	0.3097	0.1148
2000+	0.3052	0.1286

Table 8: Correlation Stratified by Set Size

The data behaved as we expected it to and there is a linear, decreasing pattern based and set size and r^2 value. However, notice that the highest standard deviation out of all of these ranges occurs in set sizes 2 to 5. We must leave open the possibility that there is simply much more spread in the group, causing an unrepresentatively high mean. Next, we wondered what happens when we further break the set sizes down by set source. Do all set sources follow the larger set/smaller correlation value trend? We chose to demonstrate this relationship graphically, which is shown below. Recall that the log of the set sizes was used instead of the actual set sizes. For this graphic particularly, the log of the set sizes provides a more accurate illustration for informal interpretation across set sources.

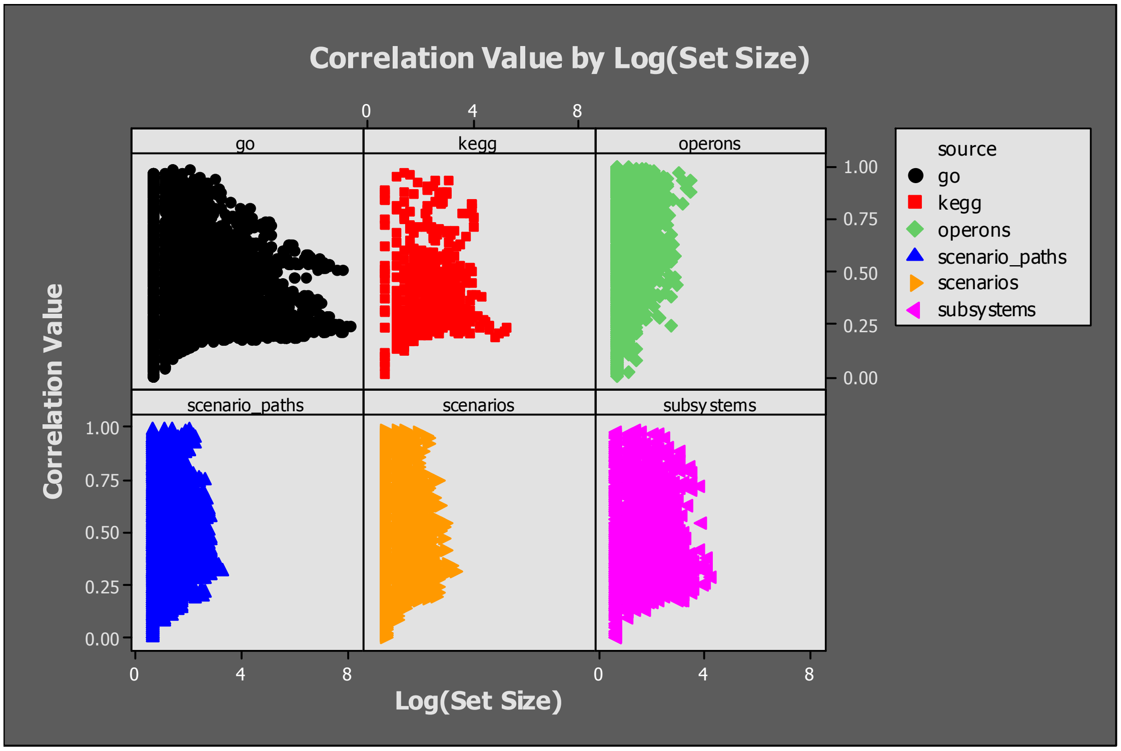


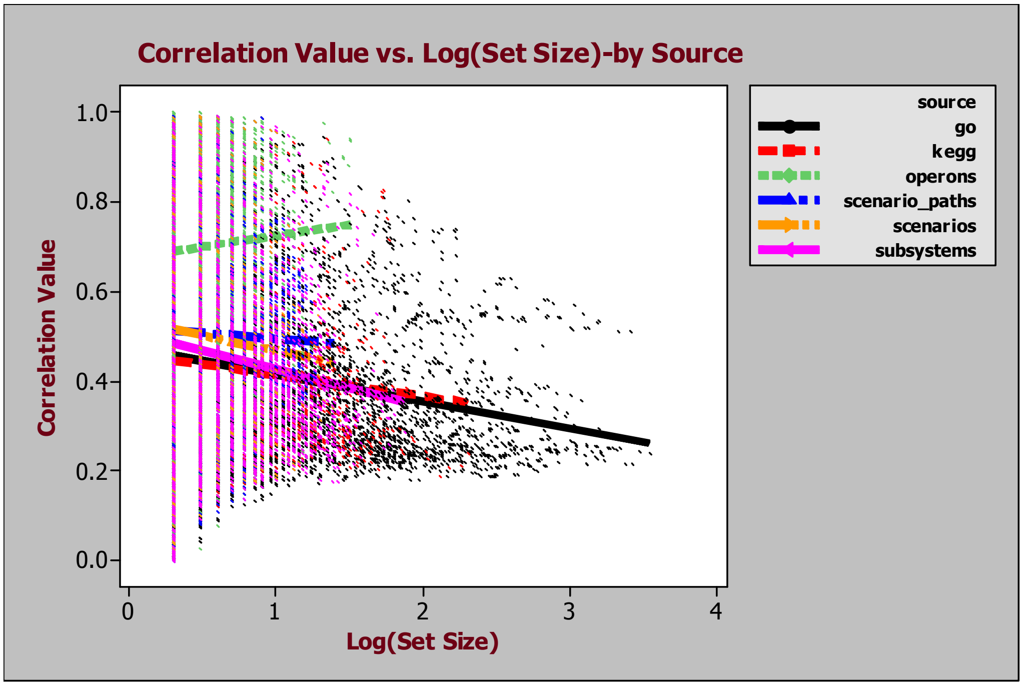
Figure 8

From this graph, it is less clear if the correlation values differ when broken down by set sizes and set source. Granted, we can see a slight downward trend for all set sources but perhaps Operons, but it is not enough to draw firm statistical conclusions. Nevertheless, it does appear that Operons remain to be the unusual case out of the set sources yet again. Interestingly enough, it appears that as set sizes increase, there may be a slight increasing pattern in the correlation values. A statistical confirmation of these visual observations follows.

V. Regression

Throughout our analyses thus far, we have seen the importance of gene set size and the effect it has on correlation. For instance, we noticed that Gene Ontology sets do not seem to be highly correlated, but also are composed of a wide variety of set sizes, many of which are massive in comparison to other set sources. We have also observed that Operons tend to have highly correlated sets, but the vast majority of these sets have between 2 and 4 genes. However, we have only informally observed this trend and have not yet statistically tested how significant this influence is. Regression is the ideal tool at our disposal to test this significance. By using multiple linear regression, we are able to see how much set size effects correlation while controlling for set source, as well as how much set source effects correlation, while controlling for set size. We can also examine the relationship between organism and set size, and the effects of any combination of set source, organism, and set size, as this interaction may also be important on correlation values. Let us reconsider a scatter plot similar to the one above, but this time on the same scale and with regression fit lines:

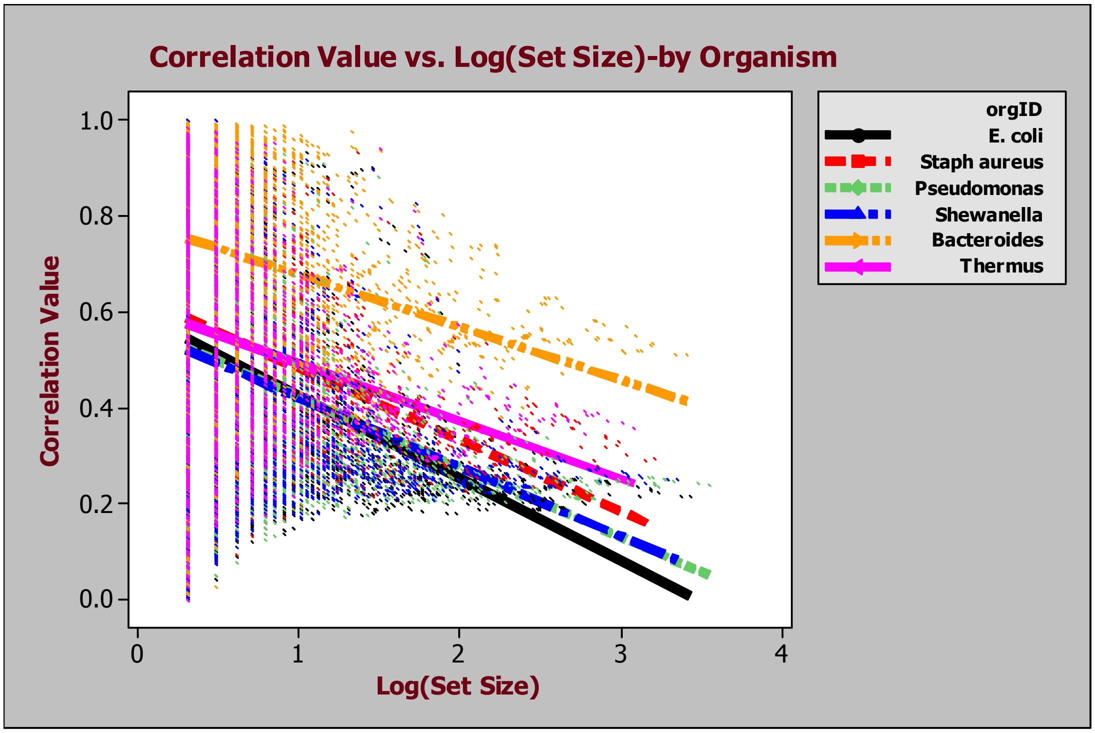
Figure 9



This scatter plot is included merely to reinforce the interesting differences between set size and correlation between set sources: Operons is the *only* set source that has a line with a positive slope. All other set sources have decreasing correlation values, as set size increases. With respect to the scatter plot above, Multiple Linear Regression can tell us 1) how different the slopes are from line-to-line (ie, set source to set source), and 2) how different the y-intercepts are from line-to-line. Note however, that since y-intercepts are where x=0, this is where log(0)=1, or where set size is equal to 1. Recall that we are only including set sizes that are greater than or equal to 2. As a result, for our purposes the y-intercept does not tell you about the correlation at the smallest set size. Instead, it may be more intuitive to think of the distance between two lines: for a fixed set size, how much more correlated is set A than set B?

We can also investigate the differences due to organism on correlation. A similar scatter plot is below, but this time broken down by organism:

Figure 10



We can see from this graph that the correlation is inversely proportional to the set size for all organisms. However, Bacteroides is noticeably higher and has a less severe slope than the other organisms. This follows the pattern that we observed earlier in our analysis.

We would now like to incorporate these factors into a multiple linear regression model. All possible factors to include in the model are size, set source, organism, size*source, size*organism, source*organism, and size*source*organism. However, before we simply throw all of these factors into our regression model, the Extra Sum of Squares F-Test[16] provides us with an elegant way to check to see if a given factor will improve the model at all and is worth incorporating. This test can compare two models at a time, given that one of the models is a special case of the other (i.e., one has an additional factor). The initial model had log(Set Size) as the only factor in a basic ANOVA test.

Below is a table of the results of this test, applied once for every additional possible factor:

Additional Factor	F-statistic	r-squared value	p-value
log(set size)	1138.01	10.3%	0.00000
source	682.33	26.5%	0.00000
size*source	10.66	26.7%	0.00000
organism	584.32	38.4%	0.00000
size*organism	15.06	38.7%	0.00000
source*organism	17.73	40.4%	0.00000
size*source*organism	3.399	40.7%	0.00000

Table 9: Extra SS F-Test Results

Based on our F-statistics, all of these additional factors could contribute significant information to the model. Notice all of the F-statistics are supported by p-values $< 1 \times 10^{-5}$. However, we should note that for some of the new models, the r-squared value does not increase by much, or the model is not covering for much more variability than is previously was. Regardless, because of our F-statistics, we will proceed with all factors, but we may ultimately conclude that they are not practically important factors and could be excluded with little harm done to the model.

Since Operons has shown to be the most unusual set source thus far by showing evidence of the highest correlation values, we choose it to be our reference group. This will help the reader to contrast set source differences. We will also choose our organism reference to be Bacteroides, since it has shown evidence of being more correlated than other organisms as well. Our model produced the following beta coefficients for the given predictors, along with their corresponding T and p-values:

β_n	Value	Predictor	T	p-value
0	0.95513	Constant	21.55	0.000
1	0.02052	Log(size)	0.35	0.729
2	-0.29511	source_go	-15.12	0.000
3	-0.30302	source_kegg	-5.23	0.000
4	-0.21120	source_paths	-6.86	0.000
5	-0.22361	source_sc	-5.85	0.000
6	-0.25228	source_ss	-8.25	0.000
7	-0.03007	size_go	-2.00	0.046
8	-0.02514	size_kegg	-1.01	0.312
9	-0.03076	size_paths	-1.46	0.145
10	-0.02422	size_sc	-0.90	0.370
11	-0.02733	size_ss	-1.41	0.158
12	-0.25434	org_E. Coli	-6.19	0.000
13	-0.20539	org_Staph aureus	-5.13	0.000
14	-0.18932	org_Pseudomonas	-4.75	0.000
15	-0.30893	org_Shewanella	-7.32	0.000
16	-0.22479	org_Thermus	-4.69	0.000
17	0.01113	size_Staph aureus	0.23	0.821
18	-0.08767	size_Pseudomonas	-1.88	0.060
19	0.02173	size_Shewanella	0.44	0.662
20	0.01029	size_Thermus	0.17	0.864
21	-0.01752	size_E. Coli	-0.36	0.719
22	0.01977	go_Staph aureus	0.52	0.607
23	-0.11092	kegg_Staph aureus	-1.30	0.195
24	-0.02061	op_Staph aureus	-0.47	0.640
25	-0.03354	paths_Staph aureus	-0.63	0.531
26	-0.01784	sc_Staph aureus	-0.28	0.781
27	-0.06514	go_Pseudomonas	-1.72	0.085
28	-0.02025	kegg_Pseudomonas	-0.24	0.813
29	-0.20792	op_Pseudomonas	-5.02	0.000
30	-0.09286	paths_Pseudomonas	-1.81	0.070
31	-0.07515	sc_Pseudomonas	-1.28	0.201
32	0.03026	go_Shewanella	0.75	0.454
33	-0.00826	kegg_Shewanella	-0.09	0.926
34	-0.01490	op_Shewanella	-0.34	0.737
35	-0.02847	paths_Shewanella	-0.51	0.609
36	0.00837	sc_Shewanella	0.13	0.898
37	0.03913	go_Bacteroides	0.89	0.372
38	0.15613	kegg_Bacteroides	1.62	0.105
39	-0.16938	op_Bacteroides	-3.61	0.000
40	0.10465	paths_Bacteroides	1.79	0.074
41	0.07313	sc_Bacteroides	1.10	0.272
42	0.09615	go_Thermus	2.06	0.039
43	-0.05616	kegg_Thermus	-0.61	0.541
44	0.00441	op_Thermus	0.08	0.933

45	-0.03127	paths_Thermus	-0.52	0.606
46	0.01068	sc_Thermus	0.15	0.877
47	-0.02617	size_go_Staph aureus	-0.60	0.550
48	0.07039	size_kegg_Staph aureus	0.87	0.383
49	-0.04236	size_op_Staph aureus	-0.62	0.538
50	0.13241	size_paths_Staph aureus	1.91	0.057
51	0.07384	size_sc_Staph aureus	0.79	0.431
52	0.07679	size_go_Pseudomonas	1.90	0.057
53	0.04208	size_kegg_Pseudomonas	0.57	0.571
54	0.20045	size_op_Pseudomonas	3.37	0.001
55	0.09501	size_paths_Pseudomonas	1.49	0.135
56	0.04944	size_sc_Pseudomonas	0.62	0.538
57	-0.01620	size_go_Shewanella	-0.37	0.713
58	0.01251	size_kegg_Shewanella	0.16	0.877
59	0.02811	size_op_Shewanella	0.44	0.661
60	0.10398	size_paths_Shewanella	1.47	0.143
61	0.00006	size_sc_Shewanella	0.00	0.999
62	-0.01149	size_go_Bacteroides	-0.23	0.819
63	-0.08141	size_kegg_Bacteroides	-0.90	0.366
64	-0.02105	size_op_Bacteroides	-0.32	0.750
65	-0.21297	size_paths_Bacteroides	-2.91	0.004
66	-0.13337	size_sc_Bacteroides	-1.43	0.153
67	-0.03060	size_go_Thermus	-0.55	0.583
68	0.09875	size_kegg_Thermus	1.08	0.278
69	-0.10611	size_op_Thermus	-1.31	0.192
70	0.05685	size_paths_Thermus	0.70	0.482
71	-0.01920	size_sc_Thermus	-0.19	0.849

Table 10: Multiple Linear Regression Results- sc=scenarios, ss=subsystems, op=operons

Here is a brief description of groupings of β values:

a) β_0: the y-intercept of the reference group, Operons for Bacteroides

b) β_1: the amount of change in correlation values of Operons for Bacteroides, due to set size.

c) β_2 through $\beta_{6:}$ the effect on correlation due to different set sources

d) β_7 through $\beta_{11:}$ the effect of set size on correlation, due to varying sources

e) β_{12} through $\beta_{16:}$ the effect on correlation, due to different organisms

f) β_{17} through $\beta_{21:}$ the effect of set size on correlation, due to different organisms

g) β_{22} through $\beta_{46:}$ the effect of set source on correlation, due to differences in organism

h) β_{47} through $\beta_{71:}$ the effect of set size on correlation, due to different set source/organism combinations

This regression model is simply too large and complex to dissect every small detail riddled within it. There are a few very important things to observe however, that give us some significant insights to the data. First note the enormous β_0 value: this confirms that Operons for Bacteroides behaved as expected in the model. Due to the small F and large p-value, β_1 tells us there does not appear to be an extremely significant effect on correlation due to set size alone for Operons for Bacteroides. We also expected this, because there is little variation in Operon set sizes. Notice all of the beta values for set source, as well as organism, are negative. This confirms all set sources and organisms are less correlated than Operons for Bacteroides. All other beta values seem to still have somewhat of a significant effect, but not nearly as strong as those due to organism and set sizes alone. As more factor interactions are added down the list, the general trend is that F-statistics get smaller and p-values get larger. We also see that the GO term is significant, but almost none of the interactions with GO are significant, which indicates GO is the reason for the poor correlation, not the other factors. We can even see that small GO and KEGG sets produce poorly correlated sets, regardless of set size.

Initially, we chose Operons as our reference set source because it has shown to be the most correlated set source so far. However, since Operons are of no interest for Pathway Analysis on multi-cellular organisms, we will choose other reference set sources as well. Particularly, we use both GO and KEGG separately as reference set sources, since they are by far most commonly used in Pathway Analysis and also have proven to produce the least correlated sets. Note that the full analysis was conducted for both of these references, but we chose to display in the table the set source predictors only, in order to focus on the important differences by set source:

Value	Predictor	T	p-value
-0.00791	source_kegg	-0.14	0.889
0.29511	source_op	15.12	0.000
0.08391	source_paths	3.02	0.003
0.07150	source_sc	2.00	0.046
0.04282	source_ss	1.55	0.12

Table 11: Regression with GO as reference set source

From this table alone, we can see that after controlling for all other factors, all other set sources except for KEGG performed better than Gene Ontology. Note however, that the T is small and the p-value is high for KEGG, indicating there may not have been a very strong difference. Now observe a similar table with KEGG as the reference:

Value	Predictor	T	p-value
0.00791	source_go	0.14	0.889
0.30302	source_op	5.23	0.000
0.09182	source_paths	1.50	0.134
0.07940	source_sc	1.22	0.224
0.05073	source_ss	0.83	0.407

Table 12: Regression with KEGG as reference set source

Similarly, we see that all other set sources performed better than KEGG, after controlling for all other factors. Once more, we have a small T and a high p-value for Gene Ontology, again indicating that Gene Ontology performed only slightly better than KEGG. Overall, our multiple regression analysis answers the question we have been hoping to answer: though set size does have somewhat of an effect on correlation, the correlation of a gene set is more heavily affected by its set source and organism than it is affected by its set size.

VI. Random Sets

One other way to compare the correlation of sets is to compare them to the correlation of random sets. If a set is highly correlated, there should be some functions or processes within the genes that correlate them in some way. Therefore, a gene set that has a random collection of genes is more likely to have genes that do not perform similar functions, so we expect a lower correlation value than a non-random set. We approached this analysis by first creating 1,000 random sets for each set size from 2-150, sampling from a list of all genes. Computation time restricted our abilities to create random sets for set sizes greater than 150, but this could certainly be executed in future analysis. Once all 149,000 random sets were created, we ran them through the same analysis as described in the Methods section to calculate the set correlation value, using the mean of all pair-wise correlation values. We then calculated what we will refer to as a "p-value" for each gene set. The formula for the p-value of a gene set A is as follows:

$$\mathrm{p-value}_{\mathrm{A}} = \frac{\mathrm{N}}{1{,}000}$$

Where N is the number of 1,000 random gene sets of the same size that had a higher correlation value than A. Obviously then, a low p-value indicates a highly correlated set and a high p-value indicates a weakly correlated set. We have included below a Cumulative Distribution Function of the p-values, broken down by set source. Since one of the questions we hope to answer through these p-values is whether set correlations are better than random, we have also included the p-values of an arbitrary random set in the CDF:

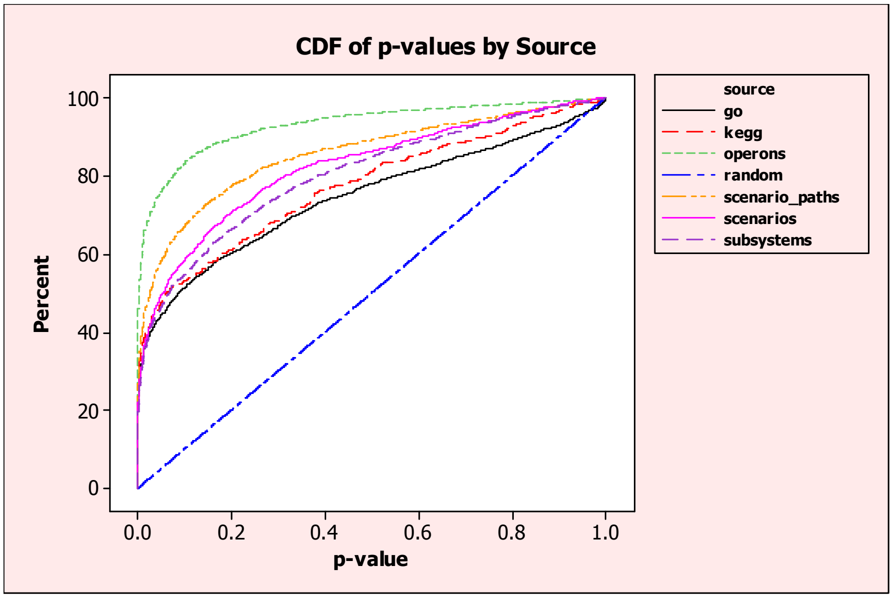


Figure 11

The first thing to note in this graph is that all sets performed significantly better than random, which is promising. As far as set source to set source correlation differences go, this graph further confirms what we have seen so far in our analysis. We again see that Operons are performing the best out of all the set sources. Nearly 80% of its gene sets have a p-value of .05 or lower, while all other set sources have 60%-80% of its gene sets with a p-value of .2 or lower. Following Operons are Paths, Scenarios, Subsystems, KEGG, and Gene Ontology, in that order. We can do a similar examination of a CDF, but this time broken down by organism:

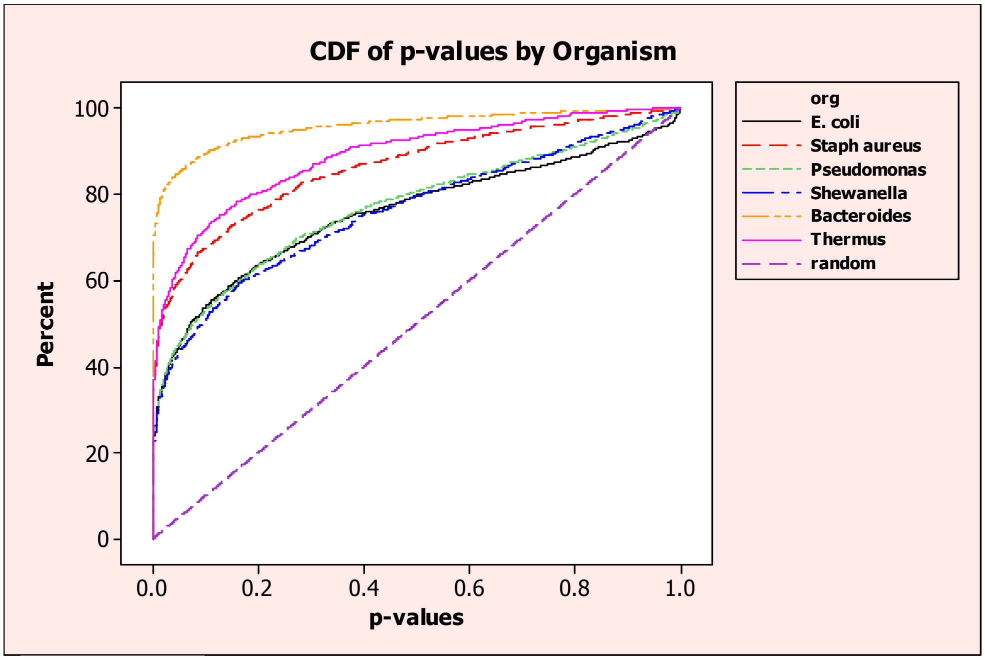


Figure 12

Once again, we see that all organisms performed better than the arbitrary random set. This graph also seems to be consistent with what we have observed thus far. Bacteroides had the highest amount of low p-values, followed by two groupings of p-value ranges of the other organisms. Again, the importance here is not necessarily which organisms are performing the best, as this is not a factor of interest in our study. Mainly, we were curious to see how consistent correlation values were across organisms overall, as this speaks of the potential for consistency of set source methods across organisms. The analysis again has demonstrated similar results to what we have seen previously: correlation is not perfectly consistent across organisms.

Conclusions

Our analysis has provided great evidence of gene set consistency differing across set sources. The importance of these gene set consistencies should not be taken lightly, as they are a key step in Pathway Analysis and Gene Set Analysis. The results have been extremely consistent across all statistical tests, graphs, and analyses conducted: Operons tend to have the most correlated sets, followed by Paths, Scenarios, and Subsystems. The least correlated sets consistently came from The Gene Ontology and KEGG. The fact that these two set sources had sets with low correlation values is a significant result, as they are the most commonly used in Gene Set Analyses.

Even though Operons, Paths, and Scenarios consistently had high set correlation values, it cannot be disregarded that the typical size of these sets were generally smaller. From our multiple linear regression model, we saw that set size does have somewhat of an effect on correlation, but not as strong as we initially expected. Most likely, their high correlations values are probably due to the way in which each of these gene sets are defined. All three of these deal with extremely specific genomic interactions and pathways, thereby limiting the amount of unrelated genes that may be in a single gene set. The most specified gene sets we looked at were those of Operons. Recall that by the mere definition of an Operon, all genes must not only be contiguous, but also be under the control of a single regulatory promoter. Operons were the only sets that applied to microbial organisms only, which helps to explain their simplistic role. It is logical then, to have obtained generally higher correlation values for these specified sets.

Though these set sources do produce highly correlated sets, there could be some potential downfalls to using them, depending on the analysis they are desired for. One of the drawbacks of these specified sets is simply that they may not apply to all Gene Set Analyses. For example, recall that Scenarios are strictly metabolic reaction networks. If a phenotype of interest in a Gene Set Analysis has nothing to do with metabolism, Scenarios are of no help. Though they have proven to be less correlated, Gene Ontology and KEGG Pathways have a wide-variety of sets, covering many different processes and functions in the genome. Gene Ontology sets should be used with particular caution however, as they proved to be erratic in their set correlations in this analysis. Though there were numerous sets with high correlation values, a typical Gene Ontology set is less correlated than desired. Ultimately, set source choice should be determined on a case-by-case basis specific to the objectives of the analysis.

Future Work

In order to better understand the organism-to-organism variation on correlation we detected, it is desirable to increase the number of organisms these set sources were collected on. We believe the specified processes of Operons, Scenarios, and Paths is the main reason for their notably high correlation values, yet we have little evidence to support this claim. In the future, we would like to reduce other sets down so they are more similar in function to these set sources to see how they perform. Mainly, we are interested in how correlated gene sets would be if they were reduced down to only genes that perform metabolic functions, so they could be better compared to Scenarios. We would also like to expand our analysis with p-values to include larger set sizes.

Acknowledgements

The early stages of this research were funded by The National Human Genome Research Institute (NHGRI) and The National Science Foundation (NSF). Thanks to Dr. Nathan Tintle, Dr. Matt DeJongh, and Dr. Aaron Best at Hope College for their collaboration on this research.

References

1. Wang K, Li M, Bucan M: **Pathway-Based Approaches for Analysis of Genomewide Association Studies.** *American Journal of Human Genetics 2007*, 81:1278-1283.

2. Chasman D: **On the Utility of Gene Set Methods in Genomewide Association Studies of Quantitative Traits.** *Genetic Epidemiology 2008*, 32:658-668.

3. Tintle N, Borchers B, Brown M, Bekmetjev A: **Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16.** *BMC Proceedings 2009*, 3(Suppl 7):S96.

4. Tintle NL, Lantieri F., Lebrec, J, Sohns, M, Ballard, D, Bickeböller, H: **Inclusion of *a priori* information in genome-wide association analysis** *Genetic Epidemiology 2009,* 33(S1):S74-S80.

5. Efron B, Tibshirani R: **On testing the significance of sets of genes.** *Annals of Applied Statistics 2007*, 1:107-129.

6. The Gene Ontology Website. http://www.geneontology.org/ . Accessed April 9, 2011.

7. The KEGG Pathway Data Base. KEGG: Kyoto Encyclopedia of Genes and Genomes Website. http://www.genome.jp/kegg/pathway.html. Accessed April 9, 2011.

8. The Known and Predicted Operons Database. Microbes Online Website. http://www.microbesonline.org/. Accessed April 9, 2011.

9. The SEED Website. http://www.theseed.org/wiki/Home_of_the_SEED. Accessed April 9, 2011.

10. The Many Microbe Microarrays Database (M^{3D}). http://m3d.bu.edu/cgi-bin/web/array/index.pl?section=home. Accessed April 14, 2011.

11. The Gene Expression Omnibus (GEO). http://www.ncbi.nlm.nih.gov/geo/. Accessed April 14, 2011.

12. Chen P, Popovich P. **Correlation: Parametric and Nonparametric Measures.** Thousand Oaks: Sage Publications Inc; 2002.

13. The R Project for Statistical Computing [Computer Program]. R; Ver: R-2.11.2.

14. The Parallel Computing Cluster at Hope College Website. http://curie.chem.hope.edu. Accessed April 9, 2011.

15. Cobb GW. **Introduction to Design and Analysis of Experiments.** New York: Key College Publishing; 1998.

16. Ramsey F, Schafer D. **The Statistical Sleuth.** Duxbury: Duxbury Press; 2002.