PREDICTING GENETIC INTERACTIONS IN THE HUMAN GENOME

By

JOSEPH KAWASH

A thesis submitted to the

Graduate School- Camden

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of Master of Science

Graduate Program in

Computational and Integrative Biology

written under the direction of

Andrey Grigoriev

and approved by

_____

Andrey Grigoriev

_____

Daniel Shain

_____

Jongmin Nam

Camden, New Jersey

May, 2013

ABSTRACT OF THESIS

Predicting genetic interactions in the mammalian genome

By JOSEPH KAWASH

Thesis Director:

Andrey Grigoriev

Genetic interaction and synthetic lethality are important tools that can be utilized to study the organization of a species genome. However genetic interaction information for mammalian and in particular human genomes is lacking when compared to other model organisms. This lack of information may be attributed to the difficulty and unreliability that seems to persist in acquiring information on genetic interactions from human cell lines. One method of resolving this problem is to use conserved genetic interactions identified in model organisms that can be extrapolated into the context of the mammalian genome. In this study, a survey is performed of genetic interaction networks from such model organisms including *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans* to test the ability of predicting genetic interactions in mammalian genomes. Additional information supporting genetic interactions, from protein interaction datasets as well as human homologs, is used to reinforce the confidence in found potential interacting gene pairs. Using orthologous human gene identifiers, networks are overlaid in order to identify potentially conserved interactions for the purpose of identifying

interacting genetic pairs in the mammalian genome. The common interactors are scored based on the model organism from which they were identified as well as their prevalence across different networks and supplemented through identification of homologous genes and human protein interactions. We find that there exist interacting gene pairs that are conserved between model organisms as well as human protein interactions. These interactions are verified using experimental information available from the literature to validate a subset of these findings.

ACKNOWLEDGEMENT

I would like to thank the members of the Grigoriev lab—Sean, Ammar, and Sulbha—for their help and support. A great amount of thanks to Dr. Grigoriev, whose expertise and guidance, as well as patience, during this project were more than I had hoped for.

And thanks to members and friends in the department, including Shirley, Rob, and Anthony, for their friendship, wit and occasional helping hand. I would especially like to thank Katarina, who was always there to support me and fill me with confidence, through failure and success.

Finally I would like to thank my family and friends, especially my parents, who have been with me and helped me through the best and worst of everything. They have been the greatest influence in my life and a constant source of support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Genetic interactions are an important element in studying the organization of a genome. To identify these interactions, pairwise combinations of two gene mutations, or deletions of genes, are observed for either the induction of cell death or a significantly measurable fitness detriment. If either of these phenotypes is observed, the gene pair is determined to be interacting (Costanzo et al, 2010). One of the subsets of a genetic interaction is cell death, termed synthetic lethality. These interactions are often found to be occurring between genes that exist either within the same pathway or functionally similar pathways (Ashworth et al, 2011). Synthetic lethal interactions are especially important when researching the organization of the genome particularly during disease, as it often identifies a specific genetic susceptibility (Farmer et al, 2005). These synthetic lethal interactions that present themselves in the genome have the possibility of being exploited through targeted therapy, unilaterally affecting a single cell line (Ashworth et al, 2011).

Genetic interaction networks have been studied extensively in several model organisms, particularly in the yeast *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Costanzo et al, 2010; Dixon et al, 2008). A study was performed using *Saccharomyces cerevisiae* to generate a genome scale interaction map through the analysis of 5.4 million pairwise interactions. The result of the study was a genetic interaction map that covered about 75% of the genome. This mapping has spawned a large amount of insight on potential genetic interactions along with the synthetic lethal subset that ultimately helps lead to a better understanding of genetic pathways and functionality in the cell (Costanzo et al, 2010).

By combining information for both synthetic lethal interactions and fitness detriments of double knockout mutants, gene interaction networks can be developed giving insight to the organization of an organism's genome that has proved useful in several model organisms. Studies have been performed using many popular model species including *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans* (Costanzo et al, 2010; McQuilton et al, 2012; Lee et al, 2008). A study has also been performed in mammals utilizing a mixed dataset from rat, mouse, dog, and human in order to attempt and better explain the organization of the human gene interaction network (Lin et al, 2010).

The human genome contains about 20,000 genes, leaving the possibilities of 200 million different gene pair combinations (Lin et al, 2010). Extracting this information can be time consuming and difficult to assess in a reasonable manner, and often times the methods that are used can be unreliable compared to the methods used in other model organisms such as yeast. In this case the human dataset was generated using a radiation hybrid approach that is much less straightforward than using the double knockout approach often found in yeast studies (Lin et al, 2010; Costanzo et al, 2010; Dixon et al, 2008). In order to alleviate some of these potential problems, it may be possible to use the available information from a compilation of these model organisms and then to extrapolate reliable genetic interactions that best relate to the human genome.

Similar attempts have been made to extract conserved interaction pairs utilizing the protein interaction models (Gandhi et al, 2006) and generating a common yeast gene interactome (Dixon et al, 2008). However these studies were limited in that they were only making observations of the protein interactions, limiting the potential for genetic research,

or in the case of yeast, did not observe the amount of conservation from species of other genera.

In the study performed by Gandhi, protein interaction datasets were utilized with roughly 70,000 interactions between the same model organisms that are planned to be used in this study (*Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans*) and were compared between all species. Few interacting protein pairs between any combination of the species were shown to be conserved relative to the amount of interacting protein pairs that were studied (Gandhi et al, 2006). The research done by Gandhi demonstrates that it is possible to identify conserved interaction pairs between several different species and shows promise that similar results can be found when utilizing genetic interactions. To support this belief, it has been shown that the protein interaction dataset does not necessarily overlap in large part with that of the genomic interaction dataset (Costanzo et al, 2010). The commonality between genetic interactions and protein interactions using the yeast dataset proves to be low, estimated between 10 and 15% (Costanzo et al, 2010). Research performed by Dixon (2008) showed that there is greater amount of overlap, almost 30%, between genetic interactions. Therefore although the combination of genetic and protein interacting pairs may indeed lead to a high confidence of conserved interaction, protein interactions alone may not be sufficient for the collection of conserved interactions between species, leaving a great amount of potentially unidentified interactions that can be supplemented through the analysis of genetic interactions.

In this study genetic interactions are surveyed across several model organisms—*Saccharomyces cerevisiae, Drosophila melanogaster*, and *Caenorhabditis elegans*—and compared with the human protein interaction database in order to

extrapolate interactions that may be conserved within the mammalian genome. Utilizing orthologs of human genes, genetic networks were created from the model organisms and overlaid to identify commonality. These common elements were then scored based on the experimental methods that were used to create their respective interaction dataset as well as the frequency that a particular pair is identified between datasets. Using this information and validating the results using human genetic interactions that were found in available literature, it was possible to identify a subset of genetic interactions that had the potential to be conserved across all species. This demonstrated the ability for conserved genetic interactions to be identified in mammalian genomes using model organisms, opening a great amount of potential in the development of targeted therapies for drug and disease study in mammals.

METHODS

Pairwise genetic interactions were collected for four sets of model organisms from various different datasets: Mammalian (Lin et al, 2010); Yeast, *Saccharomyces cerevisiae* (Costanzo et al, 2010); Fly, *Drosophila melanogaster* (McQuilton et al, 2012); and Nematode, *Caenorhabditis elegans* (Lee et al, 2008).

The largest dataset was the mammalian dataset containing about seven million gene interaction pairs. The other datasets had far fewer numbers of interactions with the yeast dataset containing 75 thousand pairs of interactors. The nematode dataset was composed of roughly 630 thousand interactions, and the fruit fly dataset yielded seven thousand gene interactions.

The reliability of the datasets was for this purpose determined by the experimental methods used to generate the dataset. In the mammalian dataset, radiation hybrids were utilized giving it the lowest amount of confidence in the interactions that were identified from this dataset. The yeast interaction dataset was generated through the analysis of double knockout strains and was considered the most robust of the datasets. Both the nematode dataset and the fly dataset were generated from curated information that is found in Wormbase and Flybase respectively and were also considered high confidence datasets.

Information from the pairwise gene interactions in yeast was graphed using python to create a gene interaction network. This graph was created with each gene representing a node and each interaction an edge. The purpose of generating this graph was to determine the betweenness and degree of each of the nodes. Fitness detriments based on double knockouts of genes (Costanzo et al, 2010) were then correlated to the average betweenness value and average degree of each pair. Other studies have identified a high amount of

correlation between either the degree of the node or the betweenness of the node and the importance that it has on the graph (Joy et al, 2004)

Pearson correlation coefficients of gene coexpression were collected using the genomes of the previously mentioned model organisms datasets and genetic coexpression information that was obtained from COXPRESdb (Obayashi et al, 2012). Pearson correlation coefficients of gene coexpression for all possible pairwise gene combinations within the genome were collected. Pearson correlation coefficients of gene coexpression using only the interacting gene pairs within the genome were also collected. These two datasets were then plotted in a histogram and overlaid for comparison.

Additional datasets were used to supplement the confidence of the genetic interactions. These datasets included human homologues collected from Homologene (Geer et al, 2010) and human protein interactions from BioGrid (Stark et al, 2006). In this way, even if a gene is not represented as a conserved genetic interaction between species, it is still represented with a moderate amount of confidence if a similar interaction exists between either human homologues or protein interactions.

A common identifier was generated for all nodes within the graphs for the purpose of comparison using Biomart (Kasprzyk, 2012). In the case of genes, human orthologs where used when available in the form of Genbank IDs. Proteins were converted to their genetic predecessor also utilizing a Genbank ID.

Interaction sets utilizing common IDs were graphed and compared using python. Pairwise interactions were tabulated referencing the dataset used to generate them. Comparisons of the output tables were then made to identify common entries between various networks.

A scoring algorithm for the genetic interactions was developed based on the prominence of a specific interaction between datasets. An interaction that is part of the mammalian dataset received one point; yeast, fly, and nematode received two points; all supplemental interaction sets were also one point. The scores of each interaction were generated by summing the value of all the datasets they are reported in. This allowed the genes that are observed in high confidence datasets repetitively to be considered likely conserved interactions and therefore received a higher score than those interactions that were present in either few datasets or an unreliable dataset. Because in this study we are looking to identify genetic interactions, the additional datasets from human protein interactions and homologs are not considered as high value as the model organism datasets.

Validation of the output was performed through a literature search of known synthetic lethal interactions in humans (Farmer et al, 2005; Bommi-Reddy et al, 2008). These interactions were collected and searched for through the outputs to validate the accuracy of the predictions made by either the individual gene interaction sets or combining the interaction datasets.

DrugBank contributed information pertaining to available drugs as well as their genetic targets where applicable. This was used in preliminary research by searching through datasets for interactions where one of the genetic partners could be influenced in some way by a particulardrug. Drug targets were identified using Genbank IDs and cross referencing these targets in the outputs generated from the networks. Tables were generated containing information on the interaction pair and the drug that would influence the pair.

RESULTS

The method of identification of genetic interactors was varied between species. It was determined that, due to the use of radiation hybrids and the large number of genetic interactions that had little validation, the mammalian genetic interaction set was not a major factor in determining a conserved or reliable genetic interaction and instead deemed a supplemental genetic interaction set.

There was no correlation identified between either the average betweenness or the average degree of an interacting pair and their respective fitness detriment posed on the cell. This was observed when analyzing the yeast genetic interaction dataset which claimed reliable and consistent fitness data for the entire interactome (Figs 1-2). This indicated that these measures, although often used in protein interaction studies as a means of determining the importance of a node or interaction set (Joy et al, 2004), may not be reliable indicators of the importance of a gene in genetic interaction studies. Although what could be identified from this information was that with the increase of either average gene betweeenness or average gene degree, it is less likely that the given interacting pair is important in the maintenance of fitness in the organism.

A key method in determining the validity of a genetic interaction dataset was the shift in the Pearson correlation coefficient of coexpression between all genetic pairs and those pairs identified as connected (Fig. 3-6). This has previously been exemplified as a reliable means of determining reliability of interacting datasets through the analysis of protein interactions in yeast (Grigoriev, 2001) However, there was no shift between these two datasets when observing the mammalian genetic interaction dataset (Fig. 6). This reinforced the notion that the mammalian dataset was not a reliable dataset to use when identifying potential genetic interactions. There was however an observable shift in the

correlation coefficients when observing the worm as well as the fly datasets (Fig. 4-5). Although there is very little shift when observing the correlation of coexpression of yeast, this was still considered a reliable dataset due to means by which the interactions were created, in this case utilizing double knockout arrays (Fig. 3).

Overall there was very little overlap between the datasets (Table 1) compared to the large number of interactions that were identified within all the organisms individually. This trend however follows other trends that were demonstrated when comparing the overlap of interactions between protein interaction data of the same set of species (Gandhi et al, 2006). In the case of genetic interactions only five were identified between yeast and nematode, and 20 identified between fly and nematode. Utilizing alternative datasets however identified more overlap. Between fly genetic interactions and human protein interactions, there were 1779 interactions identified. Utilizing the nematode dataset and human protein interactions, there were 566 interactions. Forty-nine entries existed between yeast genetic interactions and human protein interactions. When all of these datasets were overlapped, it was found that 18 entries existed between nematode, fly, and protein and only a single entry exists between nematode yeast and protein (Table 2). These overlapping interactions were considered of greater confidence than any of the interactions that were only available in a single dataset and relied simply on the identification of human orthologs for their existence.

The validation of the network using literature curated genetic interactions found four synthetic lethal genetic interactions that span multiple output datasets. The BRCA1-PARP interaction (Farmer et al, 2006) as well as VHL–MET, VHL-CDK6 and VHL–MAP2K1 (Bommi-Reddy et al, 2008) were all found in at least the fly and the worm interaction datasets. In the case of BRCA1-PARP, the interactions were found in the

overlapping datasets of nematode and fly as well. All were supplemented through the identification of human homologs as well as being present in protein interactions.

DISCUSSION

Overall there were a very high number of interactions that were generated by each of the graphs individually, especially the mammalian interaction dataset. Although not all the information in these datasets proved to be useful or reliable in the case of identifying conserved genetic interactions. The mammalian dataset was considered to be of little use because the interacting pairs were identified using radiation hybrids, and it was felt that this was not as reliable as using double knockouts such as in yeast. This is also shown through the graphs demonstrating the difference between Pearson correlation of genetic coexpression. In the mammalian dataset, there is very little difference in coexpression of all genetic pairs and interacting genetic pairs (Fig. 6). The datasets from fly and nematode however showed a much greater shift between all genetic pairs and interacting pairs (Fig. 4-5). This is a much better indication of a reliable dataset as pairs that are considered interacting should be expressed at similar time in order to compensate for the loss of a partner. The Pearson correlation coefficient of yeast did not show a great amount of shift between all pairs and connecting pairs (Fig. 3). This may be explained through the fact that the yeast interaction data also takes positive interactions into account, and although there is little shift, this is still considered a reliable dataset due to the means by which the pairwise interactions were generated. The yeast interaction set utilized a double knockout approach and then assayed the growth of the colony on media, so it is very straightforward if the interaction of two genes is required for cellular viability.

In an attempt to correlate the interaction occurring with a fitness detriment, measures of betweenness and degree of each gene were calculated from the yeast dataset

and plotted against the previously determined fitness detriment. It was determined that there was no correlation between either gene degree or gene betweenness and the amount of detriment this had to the cell, therefore this information could not be reliably used to estimate the importance of an interacting pair.

In addition to using Pearson correlation coefficient of coexpression as a means to determine the value of an interaction dataset, the datasets were overlaid with each other to find out the amount of commonality between them. Overall there was very little that was common between all of the datasets. There were no common entries between all of the genetic datasets although pairwise overlap was identified between yeast and nematode along with fly and nematode. In order to strengthen the existing overlapping results as well as provide an additional means of validating the singular results, alternative interaction datasets were used and compared. In this case the use of protein interaction information was utilized as a reliable alternative dataset. The human protein interaction dataset was used because the goal of this work is identifying conserved human genetic interactions, and due to the large amount of potential variance between species proteomes, this difference in proteomes was a variable that we did not want to consider. Overlapping human protein interactions with genetic interactions from the various model species identified far more potential interacting partners in all three species than using genetic interactions alone. In addition it was possible to identify most of the genetic interactions existing in nematode and fly that also existed in human protein interaction data (18 out of 20). This information supports the idea that genetic interactions that are conserved between model organisms are likely to be expressed in the mammalian genome.

To validate the interactions, found known synthetic lethal interactions were identified in the research. In this case interactions found were BRCA1-PARP (Farmer et al,

2006), VHL – MET, VHL - CDK6 and VHL – MAP2K1 (Bommi-Reddy et al, 2008). These interactions were identified in the output across nearly all datasets with the exception of yeast, indicating that conserved interactions from the model organisms can be extrapolated to identify synthetic lethal in mammals. Additionally, all of these interaction sets were identified in human protein interaction datasets. This suggests that the use of protein interaction information may provide a great deal of insight to determining the validity of a genetic interaction. These interactions would be conserved between other organisms and can be especially helpful when working to predict mammalian genetic interactions.

To provide additional insight to the potential these interactions have regarding the development of targeted drug therapies, information was utilized from DrugBank concerning genetic targets of several drug compounds. The genetic interactors that were previously identified in the overlapping datasets were found to often times have multiple drug compounds that would be interacting with one of the pairs of genes. This result is promising showing that it is possible to use such an analysis to unilaterally target a specific mutation within a cell line with a drug regime that can hinder the viability of the mutated cell. This can lead to a great amount of potential towards the advance of research in fields such as cancer drug development and therapy where unilaterally mutated cell lines can be identified and more specifically targeted.

REFERENCES

Ashworth, A., Lord, C.J., and Reis-Filho, J.S. (2011). Genetic interactions in cancer progression and treatment. Cell *145*, 30-38.

Bommi-Reddy, A., Almeciga, I., Sawyer, J., Geisen, C., Li, W., Harlow, E., Kaelin, W.G. Jr., Grueneberg, D.A. (2008). Kinase requirements in human cells: III. Altered kinase requirements in VHL-/- cancer cells detected in a pilot synthetic lethal screen. Proc. Natl. Acad. Sci. U.S.A *105*, 16484-16489.

Constanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science, *327*, 425-431.

Dixon, S.J., Fedyshyna, Y., Koha, J.L.Y., Prasad, T.S.K., Chahwana, C., Chuad, G., Toufighia, K., Baryshnikova, A., Haylese, J., Hoef, K., et al. (2008). Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. PNAS *43*, 16653-16658.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature *434*, 917-921.

Gandhi, T.K.B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagarahu, S., Periaswamy, B., et al. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat. Genet. *38*, 285-293.

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S.H.. (2010). The NCBI BioSystems database. Nucl. Acids Res, *38*, 492-496.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucl. Acids Res. *29*, 3513-3519.

Joy, M.P., Brock, A., Ingber, D.E, and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. J. Biomed. Biotechnol. *2005*, 96-103.

Kasprzyk, A. (2012). BioMart: driving a paradigm change in biological data management. Database *2012*, bar049.

Lee, I., Lehner, B., Crombie, C., Wang, W., Fraser, A., and Marcotte, E.M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*.

Nat. Genet. *40*, 181-188.

Lin, A., Wang, R.T., Ahn, S., Park, C.C., and Smith, D.J. (2010). A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. Genome Res. *20*, 1122-1132.

McQuilton, P., St Pierre, S.E., Thurmond, J., FlyBase Consortium. (2012). Flybase101—the basics of navigating FlyBase. Nucl. Acids Res. *40*, 706-714.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Motoike, I., and Kinoshita, K. (2013). COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. Nucl. Acids Res. *41*, D1014-D1020.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucl. Acids Res. *34*, D535–539.

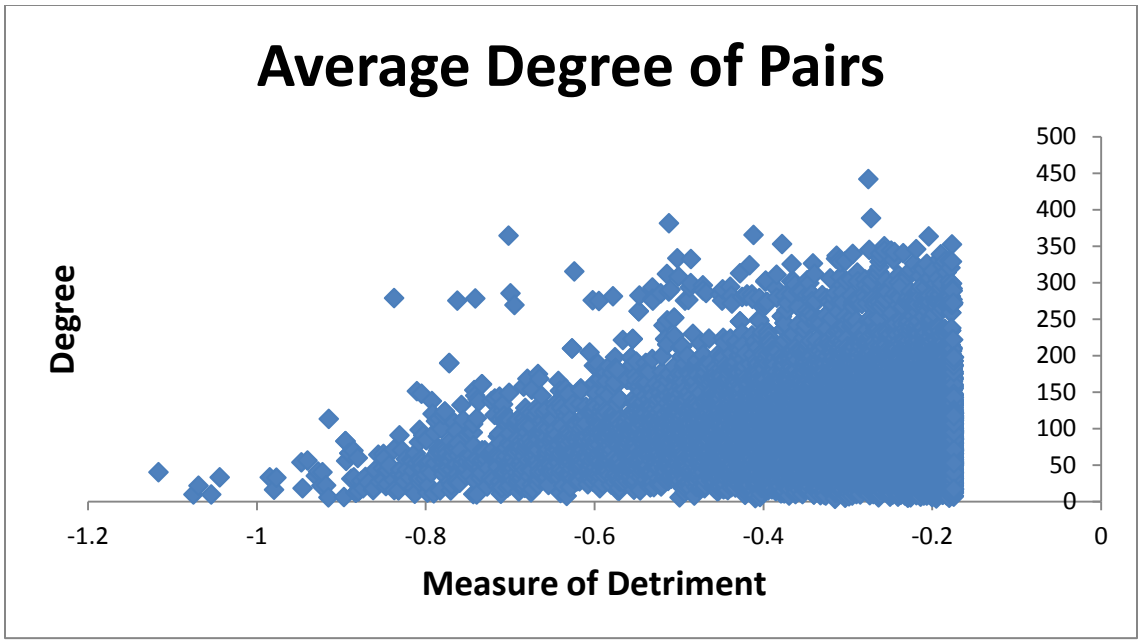

Figure 1. Degree vs. fitness detriment. Graph depicting the average degree of the two genes making up a genetic interaction compared with the measured fitness detriment of the loss of the gene pair from the yeast dataset.

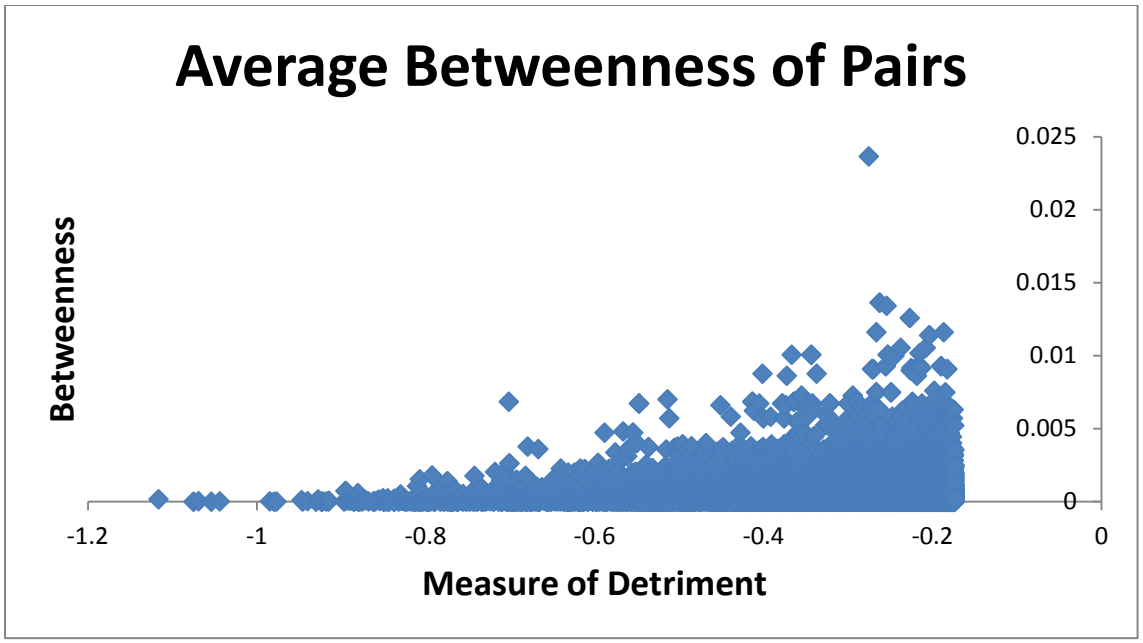

Figure 2. Average betweenness vs. fitness detriment. Graph depicting the average betweenness of the two genes making up a genetic interaction compared with the measured fitness detriment of the loss of the gene pair from the yeast dataset.

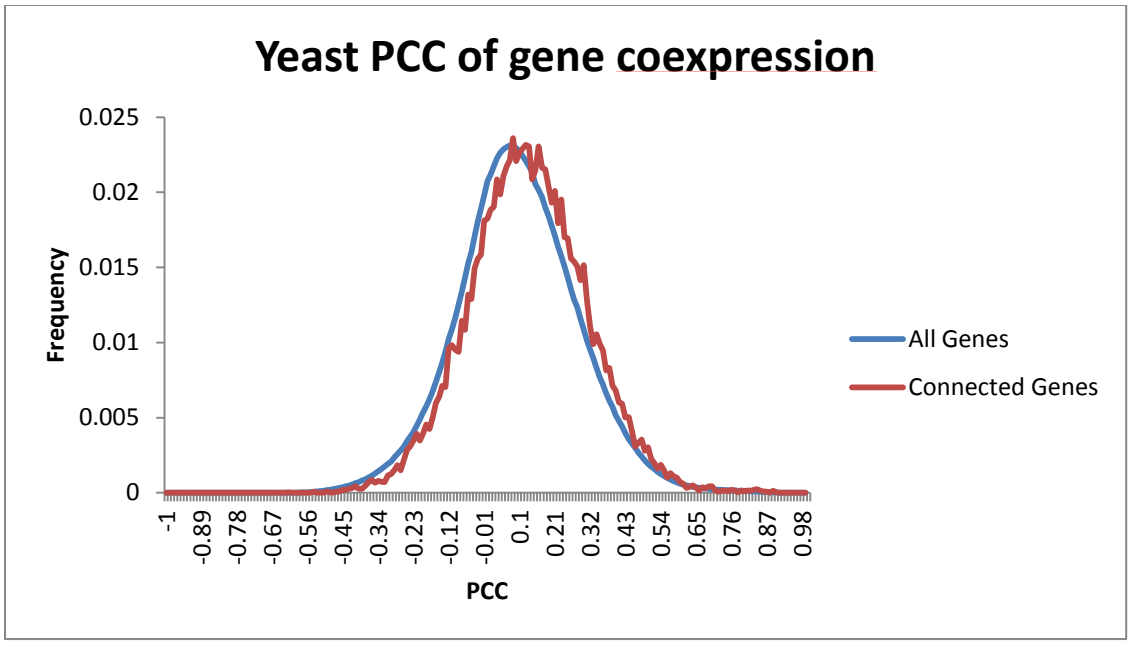

Figure 3. Pearson correlation coefficient of gene coexpression in yeast. Histogram comparing the overlap of PCC of gene coexpression between all possible gene pairs and connected gene pairs in yeast.

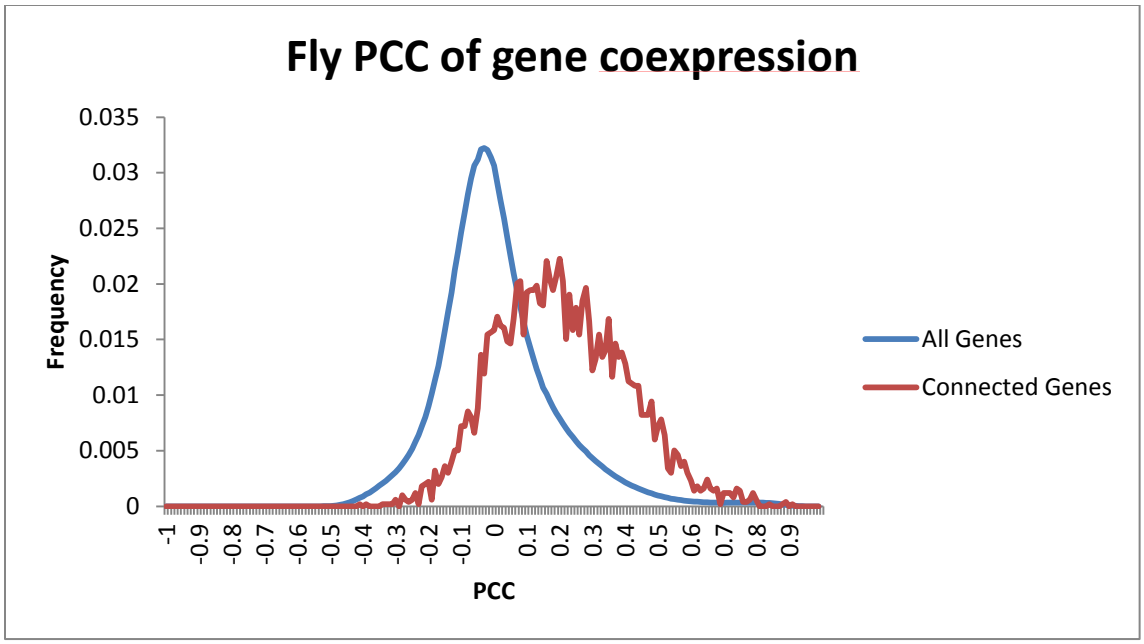

Figure 4. Pearson correlation coefficient of gene coexpression in fly. Histogram comparing the overlap of PCC of gene coexpression between all possible gene pairs and connected gene pairs in fly.

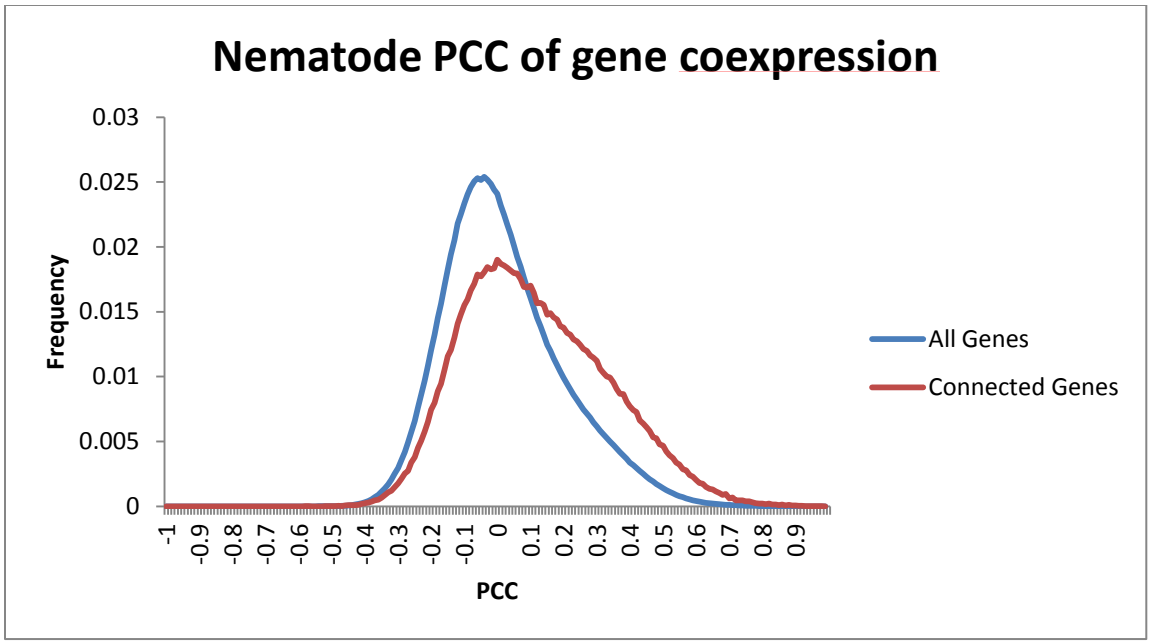

Figure 5. Pearson correlation coefficient of gene coexpression in nematode. Histogram comparing the overlap of PCC of gene coexpression between all possible gene pairs and connected gene pairs in nematode.

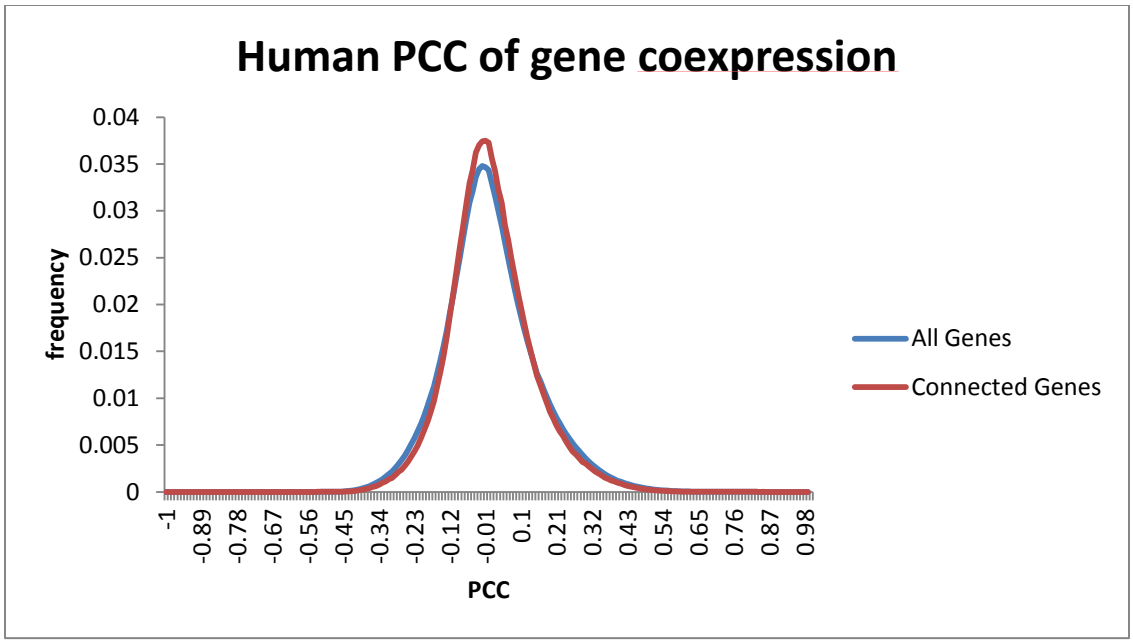

Figure 6. Pearson correlation coefficient of gene coexpression in humans. Histogram comparing the overlap of PCC of gene coexpression between all possible gene pairs and connected gene pairs in humans.

Gene	Interactant	Network Species
CH471098	BC007546	W
CH471098	D13811	W
CH471098	D13811	W
CH471098	D13811	W
CH471098	M69175	W
CH471098	M69175	W
CH471098	M69175	W
CH471098	X79193	W
CH471098	X79193	W

Figure 7. Example of output. Example of genetic interactions found in the output generated from querying human orthologs in the worm gene interaction network.

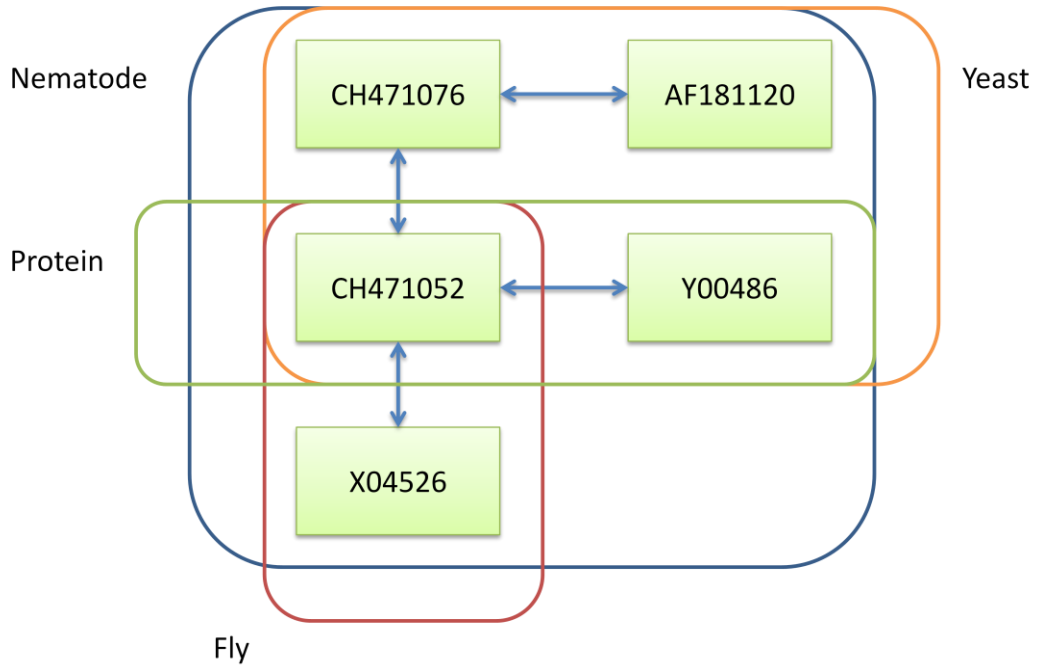

Figure 8. Example of overlapping network results. A small example of a genetic interaction network from the output using the nematode network, and the overlaps that exists between other networks utilized.

Interactions	CH471076 AF181120	CH471076 CH471052	CH471052 Y00486	CH471052 X04526
Nematode	2	2	2	2
Fly	0	0	0	2
Yeast	2	2	2	0
Protein	0	0	1	0
Total	4	4	5	4

Figure 9. Example of scoring. An example of scoring interactions present across multiple networks. Each network is assigned a value, with the sum of the values being assigned to the interaction giving a measure of its potential conservation.

	All	Mamma-Fly	Mammal-Nematode	Mammal-Yeast	Fly-Yeast	Fly-Nematode	Nematode-Yeast
Genetic Interactions	0	~	~	~	0	20	5
Protein Interactions (Gandhi 2006)	16	43	15	0	0	36	0

Table 1. Common entries between datasets. A summary of the common interactions found between genetic interaction networks and common interactions found between protein interaction networks (Gandhi et al. 2006).

	Fly	Nematode	Yeast	Nematode-Fly	Nematode-Yeast
Genetic and Human Protein Interactions	1779	20	49	18	1

Table 2. Common entries between datasets including proteins. A summary of the common entries that were found between genetic interactions and their overlap in the human protein interaction network.