Mutant vs. Gene Expression Profiles for Function Prediction

Tomaz Curk,[a] Uros Petrovic,[b] Gad Shaulsky,[c] Blaz Zupan[a,c]
a) University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
b) J. Stefan Institute, Department of Biochemistry and Molecular Biology, Ljubljana, Slovenia
c) Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX
tomaz.curk@fri.uni-lj.si, blaz.zupan@fri.uni-lj.si, gadi@bcm.tmc.edu, uros.petrovic@ijs.si

Abstract

A popular utility of microarray data is to make inference on gene function based on similarity of its expression to expressions of other, functionally already annotated genes. This approach may use available collections of gene expression measurements that study organisms under different conditions. An alternative way, enabled by recent advances in biotechnology, is to associate gene function to a phenotype of its corresponding mutant defined by expressions of all other, not mutated genes. In the paper, we use a technique called gene-coexpression networks to compare the two approaches, and apply it to data on budding yeast *S. cerevisiae*. In terms of gene function prediction and contrary to our expectations, we found that mutant phenotypes are on the overall not more informative than gene expression profiles, and we provide biological explanation why some gene functions can be better predicted using one type of data and not the other.

1 Introduction

The development of DNA microarrays allows whole-genome expression profiling, measuring the expression of each of the genes in a single assay. Changes in gene expression are often related to specific cellular needs and most expression profiling studies try to identify genes that respond to specific conditions or treatments. Recently the idea that expression of all genes could be used as an indication of cellular state has received great attention [Alizadeh *et al.*, 2000; Bittner *et al.*, 2000; Hughes *et al.*, 2000]. Whole-genome expression profiles of mutants thus hold great promise for rapid genome function analysis. It is plausible that the mutant expression profile could serve as a universal phenotype [Hughes *et al.*, 2000; Hughes, 2005; Van Driessche *et al.*, 2005] and as such is believed to be very informative for assigning gene function.

Instead of associating gene function to its expression pattern under different conditions, we can consider the entire microarray profile of a strain that is mutated in one gene as an indicator of that gene's function. This method has been demonstrated successfully in yeast [Hughes *et al.*, 2000] and in cancer cell characterization [Alizadeh *et al.*, 2000; Bittner *et al.*, 2000]. The reason to use this alternative also follows the finding that gene expression and gene function show very little correlation on a global scale (less than 10% of the cases) [Winzeler *et al.*, 1999]. When reasoning on gene function classical genetics has much depended on observational mutant phenotypes (*e.g.* "mutant grows", "does not grow", "sporelates", etc.), leading us to believe that their modern transcriptional variants will provide a robust funding for function prediction.

The study reported here was inspired by investigation of Stuart *et al.*, who showed that, in contrary to above, predicting gene function using evolutionary conservation in the wild is more sensitive than scoring the phenotype resulting from strong loss-of-function mutants in the laboratory [Stuart *et al.*, 2003]. They base their work on the assumption that genes commonly found in diverse organisms and with by-organism correlated expression patterns under a large number of diverse conditions imply functional relation. In order to distinguish accidentally co-regulated genes from those that are physiologically important they observe the evolutionary conservation between multiple species (yeast, worm, fly and human). They believe that evolutionary conservation is a powerful criterion to identify genes that are functionally important from a set of co-regulated genes and that co-regulation of a pair of genes over large evolutionary distances implies that the co-regulation confers a selective advantage, most likely because the genes are functionally related [Stuart *et al.*, 2003]. Using the method of gene-coexpression networks they were able to identify several examples of evolutionary conserved functional groups with high gene coexpression. Importantly, they also show that predictive accuracy is much poorer when using only the data on a single-organism.

In this paper we compare the level of information in mutant expression data with the information in gene expression data. Similarly to Stuart *et al.* we base our work on the assumption that similar expression profiles imply similar function and use the method of gene-coexpression networks to make the comparison.

The method of gene-coexpression networks requires a measure of distance in gene expression. The most common way, the one also used in [Stuart *et al.*, 2003], to measure distance (or similarity) in gene expression between pairs of genes is to compare their expression profiles, where expression of a gene is measured under different experimental conditions. We will call this type of

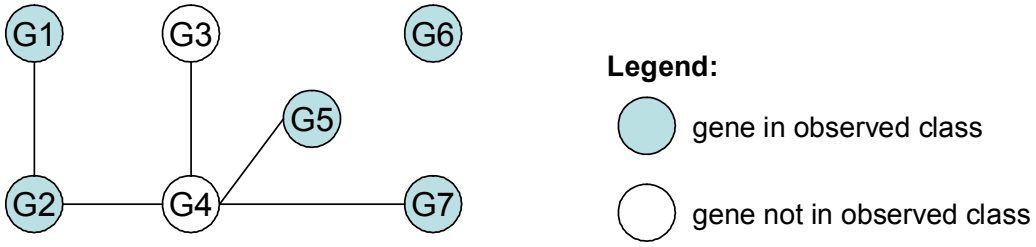


Figure 1. Gene-coexpression network for seven genes (G1-G7). We connect those pairs of genes that are correlated more than an arbitrarily selected threshold (some genes might not be connected, gene G6 in this example). There are five genes annotated to the observed class (nodes G1, G2, G5, G6 and G7), but only two of them are connected to each other (G1 and G2). The class coverage of this network is then: 2/5=0.4. To calculate accuracy we need to count edges. There is only one edge connecting two genes in observed class (edge G1-G2). There are four edges coming from class genes (edges G1-G2, G2-G4, G5-G4, G7-G4). The accuracy of this network is then 1/4=0.25.

data *gene profile*. As an alternative, we measure distance between genes using transcriptional profiles of mutant strains. For a mutated, usually deleted gene, expression of all genes in mutant's genome is measured. When comparing two genes, we compare transcriptional profiles of the respective two mutant strains. We will call this type of data *mutant profile*. In both cases and like in the study by [Stuart *et al.*, 2003], we use Pearson correlation as a distance function.

2 Data and methods

In this section we describe the microarray gene expression data sets and gene functional annotations used. We also describe the method of so called gene-coexpression networks that we applied to measure the ability to predict function from gene coexpression.

2.1 Gene expression and functional annotation

We have used two data sets of microarray gene expression measurements. For *gene profiles* we used data from a study of cell-cycle in *S. cerevisiae* where whole-genome expression under 73 conditions was measured by [Spellman *et al.*, 1998]. For *mutant profiles* the data was obtained from a compendium of whole-genome expression measurements of 300 diverse mutants and chemical treatments in *S. cerevisiae* as performed by [Hughes *et al.*, 2000].

To test the results of our predictions we have used existing functional annotations on 76 GO slim terms, which is a collection of high level Gene Ontology (GO) terms [Ashburner *et al.*, 2000] that best represent the major biological processes, functions, and cellular components that are found in *S. cerevisiae* (data available at http://www.yeastgenome.org). We also used KEGG annotation [Ogata *et al.*, 1999] for four functional classes described and used in [Stuart *et al.*, 2003]: Cell cycle, Proteasome, Oxidative phosphorylation and Ribosome.

2.2 Gene-coexpression networks

The method presented in [Stuart *et al.*, 2003] measures the correlation between gene coexpression and function. The method, called *gene-coexpression networks*, requires a measure of gene coexpression (or distance in gene expression) in order to build a network of coexpressed genes. In their paper, Stuart *et al.* used gene expression measured under different conditions. We have applied the same method to relate genes based on their mutant-based transcriptional phenotypes.

A gene-coexpression network is a graph where nodes represent genes. Edges in this network connect two nodes if coexpression of their corresponding genes is higher than an arbitrary threshold. By varying this threshold we can generate different networks: from relatively unconnected networks, where only the most coexpressed genes are linked, to highly connected networks with edges relating also genes with low correlation. Each time we can measure the connectivity properties of genes from a selected functional class. One such measure of connectedness is *coverage*: the percentage of class genes that are connected to at least one gene from same class. The other is *accuracy*: the number of edges connecting genes from same class divided by the number of all edges coming from class genes (see Figure 1 for example).

Gene-coexpression networks can be seen as a method to cluster genes. At the same time, by varying the threshold, they also give a general overview of the relation between function and gene coexpression. In their study, [Stuart *et al.*, 2003] verified the significance of the interactions in such networks by means of a variety of statistical and permutation tests. They compared the number of interactions (links) in random networks with real networks, the influence of selection of microarray experiments on the ability to identify interactions, and the influence of noise in microarray data on the constructed networks. They found the method to be robust and appropriate for the task. For details see [Stuart *et al.*, 2003]. Among other things, they showed that genes from some functional classes were highly inter-connected in the coexpression network, indicating a correlation between function and coexpression.

2.3 Performance of gene-coexpression networks

We can plot coverage and accuracy values of gene-coexpression networks obtained at different thresholds of gene coexpression for a selected functional class of genes (see Figure 2a). At high thresholds the class coverage is close to zero, because the networks include only a few edges. If genes from the same functional class are highly connected (and at the same time disconnected from genes

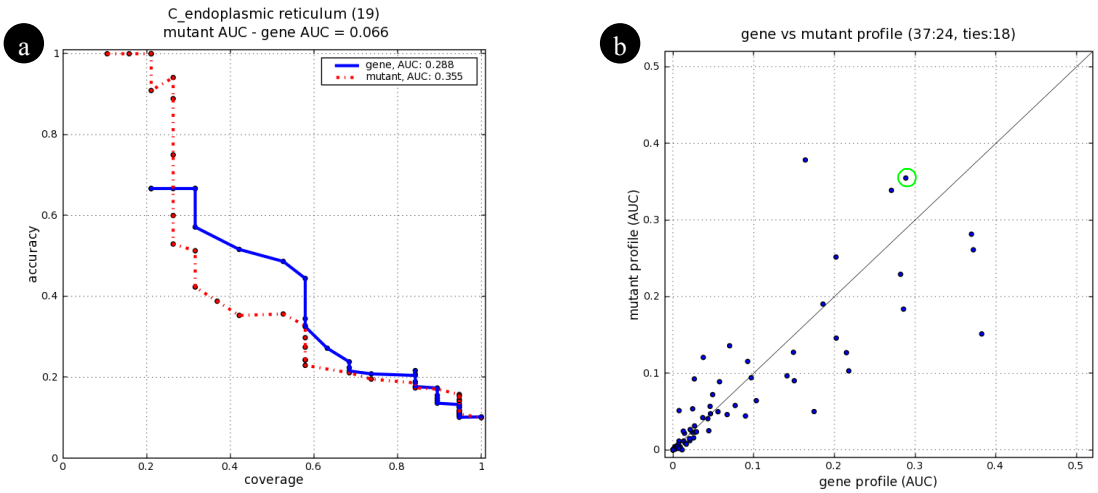


Figure 2. a) Comparison of the two performance curves of gene-coexpression networks built for class "C: endoplasmic reticulum" (C indicates that the term is from "cellular component" aspect of GO). There are 19 genes in the class (number indicated in parentheses). Solid line curve shows the performance of gene-coexpression network built using gene profiles (with AUC = 0.288), and dash-dot line curve the performance when using mutant profiles (with AUC = 0.355). A simple comparison of the two values tells us that mutant profiles are more correlated than gene profiles for the selected functional class. **b)** Graph showing AUCs obtained using gene profiles (X axis) and mutant profiles (Y axis) for ~80 functional classes. Each point represents a functional class. Its X coordinate is AUC of performance curve obtained using gene profiles, Y coordinate is AUC of performance curve obtained using mutant profiles. Encircled, at coordinates (0.288, 0.355), is class from example in Figure 2a. Gene profile wins 37 times, mutant profile wins 24 times. There are 18 ties – cases when both AUCs are equal.

in different functional class), we obtain high accuracy. Relaxing the threshold, coverage monotonically increases at increasing risk for lower accuracy (accuracy may increase, however, but would on average decrease monotonically; see mutant curve in Figure 2a). The closer is the curve to point (1, 1) – the highest coverage and the highest accuracy – the better. By calculating the *area under the curve* (AUC), we can summarize the two measures into a *single performance value* that describes the level of correlation between gene function and coexpression.

2.4 Quantitative comparison of gene function prediction from gene and mutant profiles

The "gene profile" curve in Figure 2a was obtained using coexpression of gene profiles from Spellman gene profile data. If we now derive coexpression of Hughes' mutant profiles and use it to build gene-coexpression networks, we can plot both curves and compare their AUCs for an observed functional class. Example in Figure 2a indicates that mutant profile might be more indicative for gene function "C: endoplasmic reticulum" because it has a higher AUC.

By observing performance plots of different functional classes we can count the number of times a profile type "*wins*," *i.e.* is more indicative to predicted class. Figure 2b is a summary of comparisons for all classes considered in this study. Each point represents a functional class, its X and Y coordinates are the AUCs of gene-coexpression networks obtained using gene and mutant profiles respectively. Points below the diagonal are functional classes that can be better predicted using gene profiles (AUC using gene profiles is higher than AUC using mutant profiles), whereas those above the diagonal are cases where mutant profiles are more indicative. By observing the number of points on each side of the diagonal, we can then easily see which profile type is generally more informative.

3 Results and Discussion

We measured the performance of gene-coexpression networks built from gene and mutant profiles for 80 functional classes. Results for selected functional classes are summarized in Table 3, where we give the difference in AUCs obtained using the two types of profiles. In Table 3 we subtracted the AUC obtained using gene profile from the AUC obtained using mutant profiles. A positive difference indicates that mutant profile is more informative, a negative difference that gene profile is more informative.

Looking at results in Table 3 and Figure 2b we find a slight indication that gene profiles might be more informative than mutant profiles. We base this on higher AUCs values for gene profiles (compare values in top and bottom rows in Table 3) and the prevailing number of times that gene profile wins in Figure 2b.

Analysis of functional classes, for which either gene expression profiles or mutant profiles are more informative than the other, resulted in a list of classes (Table 3) that are in accordance with current understanding of regulation of cellular functions in yeast. There are some functional classes of yeast genes that are transcriptionally regulated and thus have relatively uniform expression profiles. This is however not a general phenomenon and other functional classes include genes coding for proteins whose activities are not regulated on the level of transcription. For some of those genes, *e.g.* for those coding for

regulatory proteins that affect transcription of other genes, it is the coexpression of their target genes that can be used for functional classification, and these classes were primarily identified through mutant profiles in our study. Functional classes for which gene expression profiles are more informative than mutant profiles included genes involved in cell cycle-related processes and functions (classes "Cell cycle", "DNA metabolism") and genes for ribosomal proteins (classes "Protein biosynthesis", "Structural molecule"). These classes of genes have been previously shown as prime examples of agreement between function similarity and gene coexpression in several studies [DeRisi *et al.*, 1997; Spellman *et al.*, 1998]. Also classified as functional classes of genes for which gene coexpression correlates with functional similarity were genes involved in metabolism and transport, *i.e.* genes coding for enzymes and transporters, and genes involved in response to stress, in agreement with previous findings [Causton *et al.*, 2001; Gasch *et al.*, 2002].

On the other hand, in the group of genes for which gene coexpression is less informative than mutant profiles, based on previous knowledge about regulation of cellular functions in yeast we expected functional classes including regulatory genes. Their expression levels do not change significantly in response to perturbations, but rather they affect gene expression of their target genes. In agreement with our expectations, the list was composed of functional classes such as "Protein kinase activity", "Protein binding activity", "Transcription" and "Transcription regulator activity", "Protein modification", "Signal transducer" and "Enzyme regulator activity", that all consist of genes coding for regulatory proteins. Unexpectedly, however, there were three additional functional classes in this group that do not contain significant amount of genes known to have regulatory functions. These classes are "Lipid metabolism", "Cytoskeleton organization and biogenesis" and "Cell wall organization and biogenesis." Intriguingly, genes belonging to these classes could have roles in cellular physiology that are, from a global perspective, more important in regulatory activities than in their direct metabolic and structural functions.

3.1 Additional experimental studies

Since some of GO annotations are inferred from expression data (in particular, this includes all annotations with GO annotation evidence codes IEP and RCA), this could be the reason for higher performance of gene-coexpression networks in general. We therefore removed IEP and RCA annotations (~5% of all annotation for yeast), and thus removed about 120 genes with no annotation left. After this procedure the results changed only slightly (see Table 4), and gene profiles still appear to be more informative (compare the two graphs in first row in Table 4).

We then performed two more tests to see how the results change if we use other gene and mutant profile data (graphs in second and third row in Table 4). First, we tried to use a different set of gene profiles (second row in Table 4), and used gene profiles from the same data set as used for mutant profiling (from Hughes dataset). In this case gene profiles consisted of measurements from approx. 270 conditions (each mutant can be seen as a condition). The mutant profile data remained the same as in our first test (expression of 6316 genes in mutant's genome). In this test, mutant profiles are slightly more informative than their gene expression profiles alternative, but the later are still winning on the overall (see second row in Table 4 and compare it to first row).

In our last test we observed if there is any difference if we look at mutant data in two different ways: as a mutant profile or as a gene profile. To do so, we used only the expression of 270 mutated genes to build mutant profiles. In this case mutant profiles become more informative (see third row in Table 4 and compare it to first and second row).

mutant - gene profile performance AUC	Functional annotation
-0.1147	P: DNA metabolism
-0.1110	P: protein biosynthesis
-0.0880	Ribo
-0.0600	P: cell cycle
-0.0564	F: structural molecule activity
-0.0528	F: transporter activity
-0.0390	F: transferase activity
-0.0219	F: oxidoreductase activity
-0.0194	P: transport
-0.0191	F: hydrolase activity
…	…
0.0109	F: signal transducer activity
0.0229	P: protein modification
0.0230	F: transcription regulator activity
0.0293	P: cell wall organization and biogenesis
0.0313	P: transcription
0.0437	P: cytoskeleton organization and biogenesis
0.0498	P: lipid metabolism
0.0657	F: protein binding
0.0662	F: protein kinase activity
0.0681	Oxid

Table 3. Difference in gene-coexpression network performance when using mutant and gene profiles. Only top ten classes for each profile type are listed. Functional classes that can be better predicted using gene profiles are listed on the top, while those better predicted using mutant profiles are shown on the bottom of the list. Prefix "P" indicates the "biological process" aspect, "F" the "molecular function" aspect of GO. Classes Ribo and Oxid are taken from Stuart *et al.*

4 Conclusion

Overall, we found no clear difference between the information coming from gene and mutant profile data. On the contrary from what was our expectation (and perhaps an unstated belief of the community), there is a slight indication that gene profiles (*i.e.* observing gene expression un-

der different conditions) might, on the overall, be more correlated to gene function than mutant profiles (*i.e.* observing expression of mutants in same condition). But, when studying a particular function, there may be a clear difference between the two approaches that can be explained with existing biological knowledge. This is a clear indication that both sources of experimental data may be used in order to successfully predict gene function. We are currently investigating ways to automatically learn how to combine both profile types for better function prediction.

The principal novelty of reported work is in direct comparison of the utility of gene expression profiles and transcriptional phenotypes of mutants for gene function prediction. The two data sources were first studied together and qualitatively compared in [Hughes *et al.*, 2000], while other references on utility of mutant-based transcriptional phenotyping are at best rare. Gene expression networks and accuracy-coverage graphs, together with utility of function annotation data bases, allowed us to compare two sources quantitatively and to draw conclusions related to particular functional groups.

Acknowledgments

This work was supported in part by Program and Project grants from the Slovene Ministry of Education, Science and Sports and by a grant from the National Institute of Child Health and Human Development, P01 HD39691.

References

[Alizadeh *et al.*, 2000] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet*, et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**(6769): 503-11.

[Ashburner *et al.*, 2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1): 25-9.

[Bittner *et al.*, 2000] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon*, et al.* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**(6795): 536-40.

[Causton *et al.*, 2001] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True*, et al.* (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* **12**(2): 323-37.

[DeRisi *et al.*, 1997] J. L. Derisi, V. R. Iyer and P. O. Brown (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338): 680-6.

[Gasch *et al.*, 2002] A. P. Gasch and M. Werner-Washburne (2002). The genomics of yeast responses to environmental stress and starvation. *Funct Integr Genomics* **2**(4-5): 181-92.

[Hughes, 2005] T. R. Hughes (2005). Universal epistasis analysis. *Nat Genet* **37**(5): 457-8.

[Hughes *et al.*, 2000] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey*, et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**(1): 109-26.

[Ogata *et al.*, 1999] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**(1): 29-34.

[Spellman *et al.*, 1998] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein*, et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**(12): 3273-97.

[Stuart *et al.*, 2003] J. M. Stuart, E. Segal, D. Koller and S. K. Kim (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**(5643): 249-55.

[Van Driessche *et al.*, 2005] N. Van Driessche, J. Demsar, E. O. Booth, P. Hill, P. Juvan, B. Zupan, A. Kuspa and G. Shaulsky (2005). Epistasis analysis with global transcriptional phenotypes. *Nat Genet* **37**(5): 471-7.

[Winzeler *et al.*, 1999] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito*, et al.* (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**(5429): 901-6.

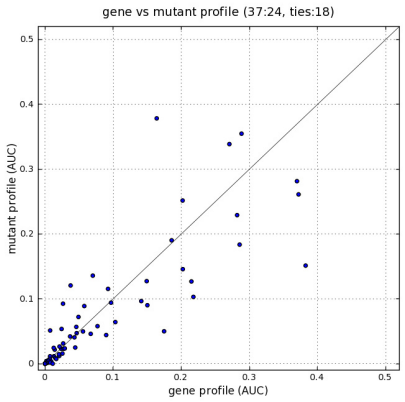
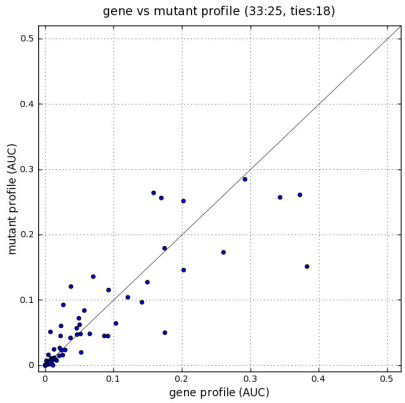
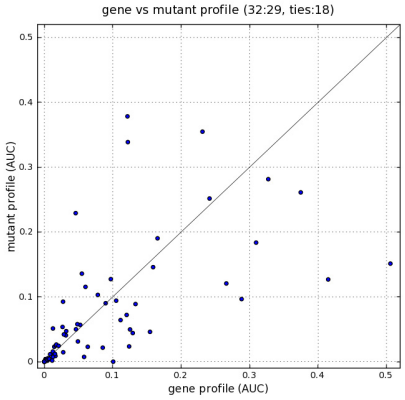
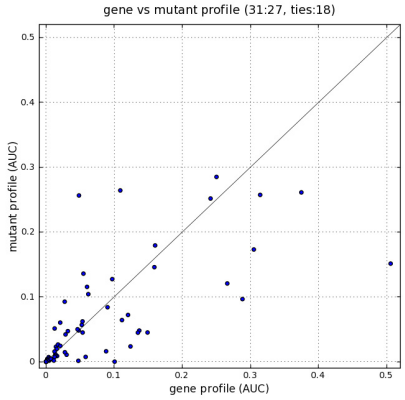
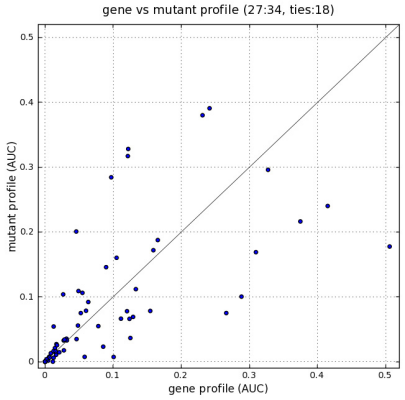
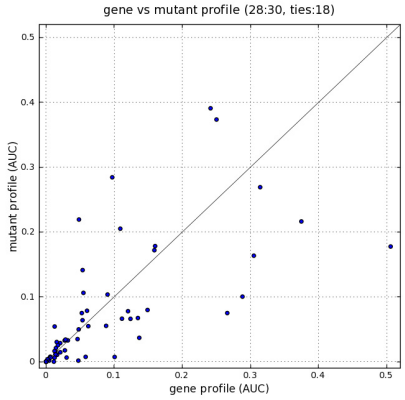
		Annotation	
		All annotation from GO slims	All annotation except IEP and RCA evidence codes (inferred from expression)
Profiles	**Gene profile:** expression in 73 conditions (Spellman data) **Mutant profile:** expression of all 6316 genes in mutant (Hughes data)		
	Gene profile: expression in 272 mutants (taken as conditions, Hughes data) **Mutant profile:** expression of all 6316 genes in mutant (Hughes data)		
	Gene profile: expression in 272 mutants (as conditions, Hughes data) **Mutant profile:** expression of only 270 mutated genes in mutant (Hughes data)		

Table 4. Comparison of performance of gene-coexpression networks built using different sources of gene and mutant profile data, for two types of annotation. Graphs in first column show the gene *vs.* mutant profile AUCs comparison when using all annotation; second column, after annotation with IEP and RCA evidence codes was removed.