BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 13 2009, pages 1694–1701
doi:10.1093/bioinformatics/btp290

Systems biology

A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide

Jonathan D. Wren*

Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation;, 825 N.E. 13th Street, Oklahoma City, Oklahoma OK 73104-5005, USA.

Received on July 7, 2008; revised on April 2, 2009; accepted on April 27, 2009

Advance Access publication May 15, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Approximately 9334 (37%) of Human genes have no publications documenting their function and, for those that are published, the number of publications per gene is highly skewed. Furthermore, for reasons not clear, the entry of new gene names into the literature has slowed in recent years. If we are to better understand human/mammalian biology and complete the catalog of human gene function, it is important to finish predicting putative functions for these genes based upon existing experimental evidence.

Results: A global meta-analysis (GMA) of all publicly available GEO two-channel human microarray datasets (3551 experiments total) was conducted to identify genes with recurrent, reproducible patterns of co-regulation across different conditions. Patterns of co-expression were divided into parallel (i.e. genes are up and down-regulated together) and anti-parallel. Several ranking methods to predict a gene's function based on its top 20 co-expressed gene pairs were compared. In the best method, 34% of predicted Gene Ontology (GO) categories matched exactly with the known GO categories for ~5000 genes analyzed versus only 3% for random gene sets. Only 2.4% of co-expressed gene pairs were found as co-occurring gene pairs in MEDLINE.

Conclusions: Via a GO enrichment analysis, genes co-expressed in parallel with the query gene were frequently associated with the same GO categories, whereas anti-parallel genes were not. Combining parallel and anti-parallel genes for analysis resulted in fewer significant GO categories, suggesting they are best analyzed separately. Expression databases contain much unexpected genetic knowledge that has not yet been reported in the literature. A total of 1642 Human genes with unknown function were differentially expressed in at least 30 experiments.

Availability: Data matrix available upon request.

Contact: jdwren@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The Human Genome Project has helped pinpoint the locations of every known human gene, but many remain functionally uncharacterized. And while some gene names may appear within one or two published papers, some of these papers are reports of high-throughput experiments and either merely mention the gene or contain only superficial details (e.g. Camargo *et al*., 2001; Lander *et al*., 2001). Some gene functions can be guessed by analysis of conserved protein domains (e.g. zinc finger domains suggest DNA binding, coiled-coil domains suggest protein–protein interactions, transmembrane domains suggest the protein is localized to a cellular or organelle membrane, etc.) and, although certainly helpful in those cases where these conserved domains are present, experimentation to test function requires knowing a bit more about the biological context. Since most genes are conditionally expressed, the biological role of a gene is strongly tied to the circumstances under which it is expressed.

Microarray technology has enabled a global view of the transcriptome and created an abundance of data. Most of the analysis tools and methods to date have focused on experiment-centric analysis. Yet, documented within this growing body of experimental microarray datasets is the behavior of individual genes responding under multiple circumstances, with strength of numbers to reinforce confidence in observed patterns of behavior. Thus, a lot of experimental data are available regarding the behavior of individual genes, gene pairs and groups of genes.

1.1 Comparability of microarray experiments

Microarrays are a powerful technology for understanding biology, and it was recognized early on that there was even more potential for discovery from combining the results of individual experiments (Khan *et al*., 1999). Combining datasets from different groups, platforms and with different normalization/pre-processing steps can be challenging (Cahan *et al*., 2007; Choi *et al*., 2007; Suarez-Farinas and Magnasco, 2007), and were first highlighted by Ghosh *et al.* (2003) who originally proposed a regression-based method (LASSO) to deal with the variation. Given the amount of noise present even in replicates of microarray experiments along with a relatively high degree of variability in individual gene expression (Pritchard *et al*., 2001; Dozmorov *et al*., 2004; Pritchard *et al*., 2006), it was not initially clear whether or not a large amount of microarray data could be combined in a meaningful way. But later studies reported that combining different gene expression datasets could yield reliable information, even those based upon different measurement technologies such as SAGE, even though there was less concordance among technologies than within technologies (Huminiecki *et al*., 2003; Jarvinen *et al*., 2004; Bammler *et al*., 2005).

*To whom correspondence should be addressed.

For analysis of multiple datasets, multi-dimensional reduction methods such as principle component analysis (PCA) and multi-dimensional data reduction (MDR) have been useful on datasets from the same platform, but when analyzing heterogeneous datasets, many experiments are missing gene-expression values, which render these methods unsuitable. So other methods of meta-analysis of microarray data have been developed, some being used to increase sample size for the study of specific diseases (Rhodes *et al*., 2002; Choi *et al*., 2004; Wang *et al*., 2004; Alexe *et al*., 2005; Yang and Sun, 2007), which tends to be heterogeneous in its clinical presentation and in strong need of cross-study comparisons (Rhodes *et al*., 2004; O'Sullivan *et al*., 2005; Fishel *et al*., 2007). Meta-analysis has also been used as a means of enhancing statistical sensitivity for determining the significance and/or robustness of expression changes (Stevens and Doerge, 2005; DeConde *et al*., 2006; Yoon *et al*., 2006; Conlon *et al*., 2007). But most of the work so far has been experiment-centric; focusing on gene sets (clusters) that recur across experiments of identical or similar types. Combining different experimental conditions would not make sense unless the focus was gene-centric—observing the co-expression patterns of all other genes while using one gene as a reference point.

1.2 Co-expression networks

Despite the technical problems inherent in combining co-expression data, several groups have recognized that this is a potentially valuable means of better understanding biology and found ways to do it. Co-expression networks have been used to visualize regulatory networks (Magwene and Kim, 2004; van Noort *et al*., 2004; Basso *et al*., 2005; Zhang and Horvath, 2005), identify co-expression modules (Yan *et al*., 2007), search for third party influences on co-expression (Li, 2002), and study the properties of these scale-free networks as a whole (Ucar *et al*., 2007; Yip and Horvath, 2007). Spellman *et al* (1998) were the among the first to use co-expression studies in Yeast, while Lee *et al*. (2004) later used co-expression studies to predict human gene function, but used Pearson's correlation coefficients to detect co-expression, which has seen use in multiple meta-analytic microarray studies (Eisen *et al*., 1998; Zhang and Horvath, 2005; Gustin *et al*., 2008; Han and Zhu, 2008). However, as shown in another study (Li *et al*., 2004) as well as here, this seems to capture only a relatively small fraction of informative co-expression patterns.

Some of the most interesting findings from these studies are that co-expressed genes tend to be conserved across evolution (Stuart *et al*., 2003; Oldham *et al*., 2006) and that the 'hubs' of these networks (i.e. genes co-expressed with many other genes) tend to evolve more slowly than the nodes on the sequence level (Jordan *et al*., 2004). These hubs, then, represent genes that contribute more towards the evolutionary fitness of an organism and alterations in their sequence or expression level are likely to be more deleterious. Yet, despite their relative biological importance, not all of these 'hub' genes have papers published describing their function.

Thus, we know that microarray experiments can be combined, at least in principle, and that their co-expression clusters correlate with gene interactions and biological function. What is not yet known is whether the patterns of co-expressed genes are important to predicting function, how many genes with no published function might be amenable to having their function predicted by gene–gene co-expression trends and how much information is derivable from high-throughput data repositories like GEO that is not present in the literature.

2 METHODS

Microarray datasets were downloaded from the Gene Expression Omnibus (GEO) repository (Barrett *et al*., 2007), which is housed at the National Center for Biotechnology Information (NCBI) on their FTP site (ftp://ftp.ncbi.nih.gov/pub/geo/). Experiments came from a total of 127 datasets (i.e. datasets describe and often contain multiple experiments). To focus specifically on the direction of the transcriptional response, human two-color microarrays were analyzed. Data were obtained from the GDS files in SOFT format. Two-color arrays were chosen for analysis to simplify the detection of relative directional change for gene–gene pairs.

Official Entrez gene names, unique gene identifiers and their associated probe (accession #) identifiers were also obtained from NCBI (ftp://ftp.ncbi.nih.gov/repository/UniGene). Only matches to primary gene names (as given by Entrez) were considered to reduce ambiguity. Among the 25 183 primary names in Entrez, only three were not unique (MAG, PTPRV and PRG2). Not all probe names were mapped to genes, but analysis of the failed mappings showed that the most frequent were control features (e.g. 3XSSC, Salmon sperm DNA, etc). Each microarray experiment was processed to identify differentially expressed genes for the meta-analysis. Since much of the information for pre-processing of the deposited raw microarray data was not available, dataset processing was done with as few assumptions as possible. Normalization was limited to regressing the mean of all expression values to zero for each array and an adaptive fold-change cutoff threshold was employed to reduce the experimental variability (Mariani *et al*., 2003). A previous study of reported (non-microarray) fold changes in the literature found that less than 5% of reported fold-changes were less than 2-fold, suggesting that this was a reasonable starting threshold (Wren and Conway, 2006). But if the total fraction of responders per experiment was above 5%, the 2-fold threshold was increased until the number of responders was $\leq$5% of all genes on the microarray, the goal of this arbitrary adaptive threshold being to increase stringency commensurate with noise level. Out of 3551 total microarray experiments processed, a total of 23 were discarded if they contained a high degree of variability/noise (i.e. $>$50% of the genes classified as differentially expressed once a 6-fold cutoff threshold has been reached).

Determining the direction co-expression patterns for gene pairs was done by first calculating the fraction of parallel patterns ($P_1P_2 + N_1N_2$/total) and anti-parallel patterns ($P_1N_2 + N_1P_2$/total), and then taking the greater of the two, where P is the positive fold change and N the negative fold change for genes 1 and 2 in any given pair of genes detected as differentially expressed. This fraction determines whether a gene–gene co-expression pair has a parallel or anti-parallel pattern of co-expression (when $>$50%) and also the 'purity' of either pattern. For example, if two genes are upregulated together in 99% of the experiments examined and in the remaining 1% one gene is up while the other is down, then the co-expression pattern is parallel and the purity is 99%. Note that there are actually four different positive–negative patterns of behavior, and detection of a parallel pattern may come entirely from P_1P_2 or N_1N_2 pairs. Thus, this is an oversimplification of all possible behavioral patterns that might be observed, but the goal here is to see if there is a functional difference between genes that respond in similar (either PP or NN) versus opposing (either PN or NP) directions.

Although the normalization method does not take into account several factors (e.g. pin printing variability, physical microarray blemishes, etc.), the large sample size and diversity of platforms minimizes the probability that false–positive or false–negative effects from one microarray experiment, series or even platform will significantly affect the overall results, except in cases of rare or weak co-regulation.

For literature studies, a software package called IRIDESCENT (Wren, 2004; Wren *et al*., 2004; Wren and Garner, 2004) was used. Briefly, IRIDESCENT uses a term thesaurus to recognize when biological 'objects'

(e.g. genes, diseases, phenotypes, chemicals, etc.) occur in text. The thesaurus is constructed using popular and freely available sources for object names (e.g. Entrez gene, OMIM, ChemID database, etc.) (Wren *et al*., 2004). For each acronym in the database, ambiguity is assessed using an acronym resolution routine (Wren and Garner, 2002) and only acronyms with less than 5% potential confusion rate are used verbatim without requiring acronym resolution. Spelling variation is partially accounted for by aligning definitions of identical acronyms (e.g. IL1 and IL-1). IRIDESCENT processes all MEDLINE records to derive an object–object co-occurrence database, which includes gene–gene co-occurrences. At the time of this study, the database was constructed using 18 438 436 MEDLINE records, 10 096 105 (55%) of which had abstracts.

Gene ontology (GO) records for the GO enrichment analysis were downloaded from the GO website in April 2008, and GO enrichment tests were performed using a chi-square test of significance. To calculate the distance between a query gene's GO categories and GO categories detected as statistically significant for a set of genes, the minimum number of nodes (i.e. categories) in the acyclic GO tree that had to be traversed to get from each statistically significant node to any of the query gene nodes (not including their inherited parent nodes) was calculated.

2.1 Calculating a mutual information measure (MIM)

The MIM takes into account the frequency of A–B co-expression relative to their respective probabilities of individual expression as follows:

$$MIM(DE) = \log\left(\frac{p(A,B)}{p(A)*p(B)}\right) \quad (1)$$

$$MIM(para) = \log\left(\frac{\left(p(Au,Bu)/p(Au)*p(Bu)\right)+\left(p(Ad,Bd)/p(Ad)*p(Bd)\right)}{2}\right) \quad (2)$$

$$MIM(anti) = \log\left(\frac{\left(p(Au,Bd)/p(Au)*p(Bd)\right)+\left(p(Ad,Bu)/p(Ad)*p(Bu)\right)}{2}\right) \quad (3)$$

In Equation (1), A and B are the probabilities of differential expression (DE) for two genes. Alternatively, a MIM can be calculated for parallel or anti-parallel patterns. In Equations (2) and (3), the subscripts u and d represent the direction (up or down) of the expression of A or B.

2.2 Ranking co-expressed genes to identify functionally similar groups

Genes were queried to select those with at least one annotated GO category. This list was then narrowed to the query genes with at least 20 genes co-expressed with them in 40 or more experiments, in both parallel and anti-parallel patterns. It is possible that genes with only parallel or only anti-parallel partners might have different properties than those with both, but this was necessary to ensure equal numbers for analysis as well as a balance of 10 parallel and 10 anti-patterns to test the effects of combining genes with different response directions. Each batch of 20 genes co-expressed with the query gene was then ranked using several different metrics. For each co-expressed gene, a measure of the 'purity' of its direction relative to the query gene was calculated (see 'Methods' section) and for every batch sent for GO enrichment analysis (Stuart *et al*., 2003; Gustin *et al*., 2008), another batch of 20 genes selected randomly was also sent. Only GO enrichment was examined (i.e. not depletion) using a chi-square test of significance (Rivals *et al*., 2007). A threshold was set to only report enrichment when four or more of the 20 genes belonged to the same GO category with $P < 0.0001$. The results are shown in Table 1.

Table 1. Comparison of different methods for prioritizing the top co-expressed genes to be sent for GO enrichment analysis and how reflective those genes are of the query gene's function

Method	Parallel	Anti	Both	Avg. Sig.	# pred/# anal	% dir hits
R^2	21 271	17 481	10 998	22.7	3748/4321 (87%)	34
Total	41 157	38 315	31 437	19.3	5012/5043 (99%)	24
MIM	12 449	13 275	6 887	13.9	4131/5043 (82%)	33
dMIM	17 281	17 045	9 822	18.1	4221/5035 (84%)	34
Purity	18 366	20 596	12 269	19.2	4451/5043 (88%)	29
T*P	42 278	39 098	29 982	22.6	4995/5043 (99%)	25
dM*P	18 691	17 319	10 192	18.9	4286/5035 (85%)	35
T*P^2*dM	31 265	24 635	18 462	25.0	4833/5035 (96%)	32

T, Total; P, Purity; dM, directional mutual information measure (dMIM). Avg. Sig., the average P-value of significant GO category enrichments, expressed as –log(P-value). On average, 12.6 ± 0.6% of random gene batches sent for analysis yielded at least one prediction. Average significance scores for random gene batches was 7.88 ± 0.37. #pred/#anal is the ratio of significant GO category enrichment (functional predictions) to the number of genes sent for analysis. The number of gene batches sent for analysis are not necessarily equal due to the constraints of ensuring a balanced selection of genes meeting minimal criteria for analysis. %dir hits is the number of significantly enriched GO categories for the 20 genes analyzed that were identical to at least one of the query gene's GO categories (top level domains not included).

3 RESULTS

A total of 3551 microarray experiments were analyzed and of the 24 553 Human genes in the Entrez Gene database, 18 516 of them (75%) were co-expressed with another gene in at least two experiments. As observed in other studies, the distribution in the number of co-expressed (CoX) genes followed a scale-free distribution, as did the differential expression (DE) for individual genes, and the two were correlated (Fig. S1, Supplementary Material).

Gustin et al. (2008) recently proposed using Pearson's coefficient to further refine the patterns of gene–gene co-expression as positive, negative or a balance of both depending upon sign. Although examples of these linear trends can certainly be seen in the data (Fig. 1), genes are also conditionally co-regulated. Thus, for genes under the same regulators we would expect that their expression levels should vary together in a linear manner, although not necessarily at equal expression levels, which can be detected by calculating Pearson's R^2 for all experiments in which the two genes were present. But for two genes whose regulatory elements differ, we would expect the R^2 value to be less sensitive in detecting co-regulation as the overlap in their conditional regulatory elements decreases. Conditional co-regulation, however, might be detectable by employing a metric from Signal Theory, the mutual information measure (MIM).

To see if MIM might be suitable to detect instances where conditional co-regulation was taking place, first random gene sets were identified where two genes (A and B) were detected as co-expressed. Since the genes were selected only on the basis of exceeding an expression threshold, the set is expected to contain a mixture of genes that are both globally and conditionally co-expressed. For genes that are co-expressed 100% of the time, it is expected that they should have relatively high MIM and R^2. For genes that are never co-expressed together, it is expected they should have very low scores in both categories. The analysis should tell us how rapidly both scores move from their highest to their

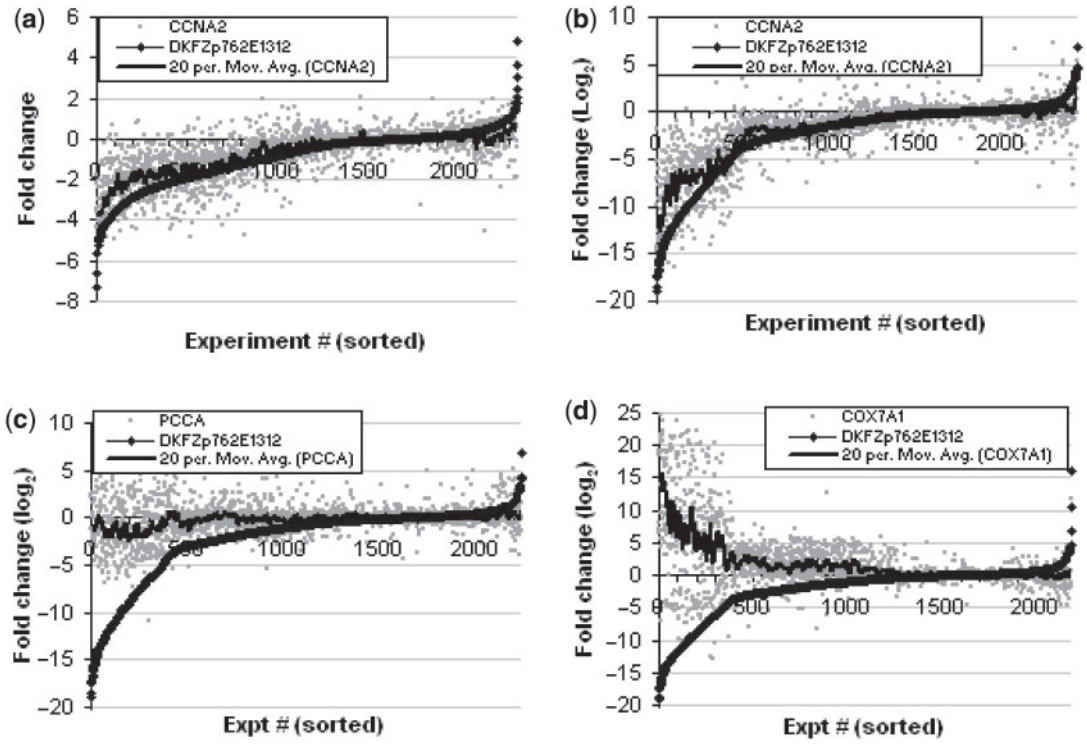

Fig. 1. Co-expression patterns for the gene DKFZp762E1312 (DKFZ, thick blue line), with values sorted from lowest fold-change (left) to highest (right). (**a**) Raw data from GEO files for experiments where DKFZ and CCNA2 are both expressed ($R^2 = 0.56$). (**b**) Same data, but normalized to the mean intensity of all values and converted to $\log_2$ value ($R^2 = 0.70$). The curve does not appear smooth because some experiments reported values as 'log ratio' but did not specify the base. The default in these cases is to assume $\log_{10}$, but the value distributions suggest some were $\log_2$ instead. Note that this affects accuracy in estimating the magnitude of the response, but not the *direction* or detection of co-expression. (**c**) Random genes were examined to ensure these patterns were not an artifact of normalization or data processing. Shown is an example of one of the random genes, PCCA ($R^2 = 0.08$). (d) Normalized data for DKFZ and COX7A1, showing an anti pattern of co-expression (Pearson's $= -0.55$, $R^2 = 0.30$).

lowest levels as gene pairs become progressively less linked in their frequency of co-expression.

A random sampling of genes (A) was taken and an average MIM and R^2 was calculated for each of the co-expressed genes (B1..n). Each metric was normalized so that it could be expressed in terms of the relative overlap of B with respect to A (i.e. the number of times B was differentially expressed when A was) and compared. A total of 98 949 co-expression samples were taken, 1000 for each percentile of overlap (the higher percentile overlaps, however, were less represented than the lower in the dataset and not all percentiles had 1000 examples even after all genes were processed). Figure 2 shows that many genes have a high degree of mutual information even with their R^2 is low. This suggests that conditional co-regulation is occurring and present within the datasets, and that estimating co-regulation based upon an R^2 cutoff may miss this subset of conditionally co-expressed genes in heterogeneous datasets.

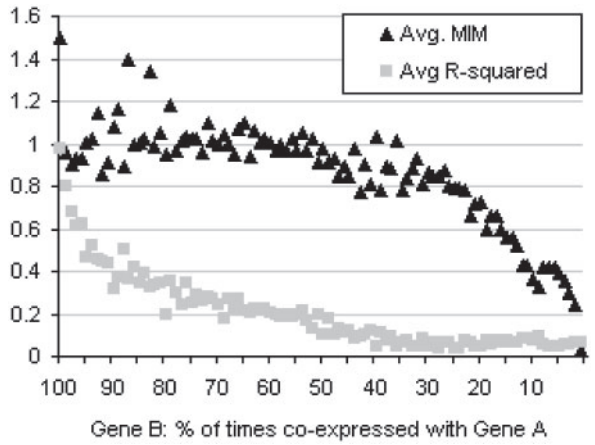

Fig. 2. Comparison of mutual information measures (MIM) versus R^2 for co-expressed genes to see if informative patterns are present even after linear correlations decline. When gene pairs are co-expressed 100% of the time, R^2 values are highest, as expected, and as the fraction drops, so does the average R^2. MIM, however, remains at relatively high levels longer than R^2.

3.1 Selecting a subset of genes co-expressed specifically and consistently with a query gene correlates with gene function

For each DE gene, there are many other genes co-expressed with it under different circumstances, with differing frequencies and patterns of co-expression. It is hypothesized that for most genes, their co-expressed genes whose 'behavior' across many different experimental conditions is *most* similar to their own would be involved in the same functions, processes and/or cellular

components. Several different methods of ranking gene sets were compared to identify which metric best corresponds to this similar behavior.

There are several ways success could be measured. The first and perhaps strongest would be the fraction of co-expressed genes belonging to the same category as the query gene beyond the number of random sets that do. This would confirm how reasonable the guilt-by-association assumption is in predicting function. However, since GO annotation is perpetually a work-in-progress in keeping up with the literature, which itself is far from complete, we could not necessarily say the genes with different categories were false-positives (e.g. they may just not be annotated). So another metric would be the number of significant GO categories found over the number predicted from random sets, since this would reflect how much functional information is returned—the more, the better. A third could be the average (or median) statistical strength of categories predicted, since this would reflect the efficiency of prioritization methods in grouping genes by their similarity. It is not obvious how much importance should be placed on each of these three means of evaluation, but in terms of predicting function, the criteria are listed in order of their relative importance.

Table 1 shows that genes ranked by R^2 receive some of the higher average statistical scores from the GO enrichment analysis, but have a lower number of predictions as well as fewer gene batches that meet minimal criteria for analysis (≥ 40 co-expressed genes with $R^2 > 0.01$). Calculating mutual information measure (MIM) of differential expression performed worse by all measures than calculating MIM of directional expression (dMIM), which is consistent with the loss of information apparent in each ranking scheme when parallel and anti patterns are combined, and consistent with the improvement seen when incorporating 'purity' into the score. Interestingly, just ranking genes by the total number of times they are co-expressed with the query gene yielded the most significant predictions. However, it also yielded the lowest fraction of direct hits.

In terms of single metrics, prioritizing by R^2 is most successful at identifying gene sets with the most functionally overlapping categories, but as shown in the table and also suggested by Figure 2, it is less sensitive at detecting functionally related groups when co-expression is conditional (i.e. fewer gene sets had minimal R^2 scores for analysis). dMIM alone enables analysis of more genes, but yields fewer total enriched GO categories. After noticing that different metrics tended to have different strengths and weaknesses, this motivated exploration of combinations to see if the strengths of one could offset the weaknesses of another. A combination of the total number of observed co-expression instances, purity of co-expression patterns and dMIM seems to provide the best balance between the three measures of success for functional prediction (i.e. relatively good functional overlap of co-expressed genes with the query gene, and more total significant GO categories returned with higher average P-values).

To control for the possibility that R^2 might be high due to highly similar experiments in each dataset, the analysis was also performed only using the first two experiments in each dataset. The results were not significantly different (data not shown).

3.2 Directionality of co-expression impacts functional predictions

In Table 1, the fraction of 'direct hits' was calculated using only the parallel genes. This is because in all the scoring schemes examined,

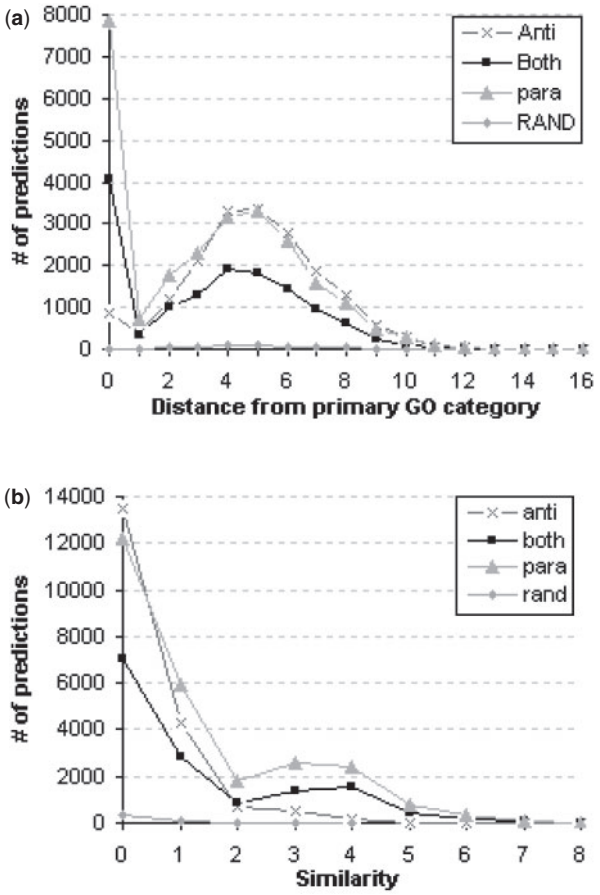

Fig. 3. Validating the assumption that consistently co-expressed genes tend to have similar functions using a GO enrichment analysis. Distance is calculated as the minimum number of GO categories that need to be traversed to get from each significant GO category ($P < 0.0001$) associated with the 20 co-expressed genes to the nearest GO category associated with the query gene. Zero distance indicates identical categories. (**a**) Distribution of significant GO categories for co-expressed genes by their distance from the nearest GO category of the query gene. (**b**) Similarity of co-expressed gene categories to nearest query gene category.

the parallel genes' functional category correlated with the query gene's functional category far more often than the anti-parallel. Shown in Figure 3a is the distribution in distances between the co-expressed gene's GO categories in the acyclic hierarchical GO tree to the nearest GO category for the query gene. That is, it shows how many nodes must be traversed to go from the GO category of the co-expressed genes to the nearest GO category of the query gene. In this figure, the ranking method (total * purity^2 * dMIM) that offered a relatively balanced performance among the criteria mentioned was used. Because the GO tree is not of uniform breadth or depth, distances are unequally distributed, as can be seen in the graph. To reduce this effect, we also used an information-theoretic method of measuring the similarity of each GO category (Lord *et al*, 2003; Resnik, 1995) (Fig. 3b). Here, we see that the parallel genes tend to have higher similarity in their GO categories than the anti-parallel genes.

Figure 3 shows that genes expressed in parallel tend to be closer in direct function as compared to anti-expressed genes, and that

lumping both anti and parallel categories together tends to produce slightly fewer statistically relevant GO commonalities than either one alone.

3.3 One example of correctly predicting unknown gene function

Figure 3 shows the efficacy of the method in general, but also included in the Supplementary Material is an analysis of 10 genes (five known and five uncharacterized), and here a specific example will be described in detail. Initially, the most frequently differentially expressed gene without any publications was DKFZp762E1312 (382 experiments). The top 20 genes co-expressed in parallel with DKFZp762E1312 (DKFZ hereafter for short) suggested it was involved in the cell cycle, by GO enrichment analysis. Among the most significant GO associations were mitosis (13 genes, $P = 4.5 \times 10^{-127}$), regulation of cell cycle (13 genes, $P = 1.2 \times 10^{-57}$) and negative regulation of DNA replication (seven genes, $P = 2.2 \times 10^{-79}$).

Figure 1a–c shows some of the global co-expression trends for DKFZ and a control (Fig. 1d). The data are noisy, but improved by normalization and co-regulation is evident for these genes as judged by both MIM and R^2 values. Involvement of of DKFZ in the cell cycle is independently suggested by cross-referencing with Cyclebase (http://www.cyclebase.org) (Gauthier *et al*., 2008), a database of time-series cell-cycle microarray experiments (Fig. S2, Supplementary Material). DKFZ is cyclically expressed during the cell cycle, peaking at the G2–M transition and bottoming out at the G1–S transition. (see also the example of PRR11 in the Supplementary Material).

Not too long after the initial analysis, two groups independently confirmed the involvement of DKFZ in the cell cycle (Kato *et al*., 2007; Luhn *et al*., 2007). It has since been renamed Holiday-junction recognizing protein (HJURP).

3.4 The data-literature information divide

The IRIDESCENT software was used (see 'Methods' section) to gain a rough approximation of the number and extent of genes characterized in the literature. This is done with the caveat that some genes may have function documented under a name not listed among the official synonyms. However, the converse is also true—a gene name may be mentioned in an abstract without any function described at all. Thus, this is an imperfect means of answering this question, but should provide a reliable approximation.

Figure 4 graphically summarizes the results—existing studies are heavily skewed towards a small fraction of the total genes (e.g. TNF-alpha, insulin, angiotensin, IL-2, etc.) and ∼37% of human genes have yet to be mentioned in a MEDLINE abstract. Although some of these genes may be mentioned within the full-text, their absence from the abstracts suggests they have not played a prominent role in the studies conducted to date. These results are not that surprising when considering that Yeast, one of the best studied experimental organisms, still has an estimated 21% of its genes uncharacterized as of 20 March 2007 (Pena-Castillo and Hughes, 2007).

For these unknown and uncharacterized genes, co-regulated genes might provide the best indication of their function, especially in the absence of informative structural or sequence-based predictive methods (e.g. conserved functional domains or TF-binding sites). To get an idea of how much information might be available from the GMA to contribute towards characterizing gene function overall, the number of times a gene was differentially expressed (DE) was plotted and contrasted with the amount of literature available for the same gene (Fig. 5). There is a correlation between the frequency of DE and frequency of publication. The most interesting feature of this analysis is that there is a set of genes for which there are no papers (marked by arrow 'a'), yet many of these unknown or poorly known genes have sufficient co-expression data to predict their function. There is also a group of genes for which no papers and no expression data exists (arrow 'b')—sort of an 'information desert' whereby neither literature nor co-expression data are available.

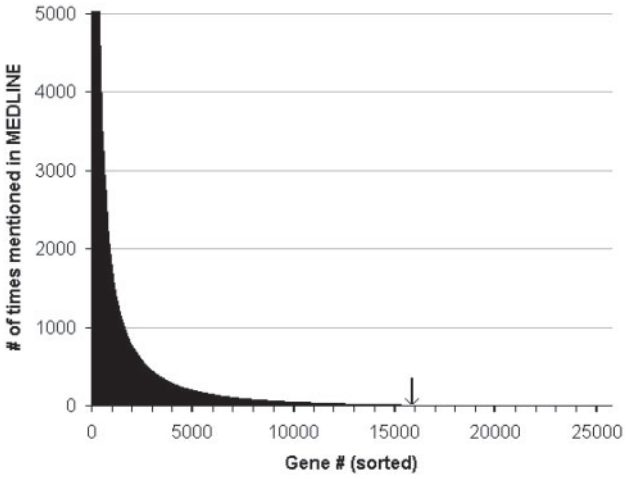

Fig. 4. An approximation of the number of times each human gene (including synonyms) has been mentioned in MEDLINE abstracts or titles. The graph is truncated on the *y*-axis at 5000 to prevent distortion (393 genes were above this cutoff). Out of 25 311 Human gene name searches, 15 977 (63%) were found at least once (gene counts differ slightly from the number of Entrez genes because the searched database not only includes miRNAs as human genes, but also incorporates HGNC and GDB human gene names). The arrow marks where the number of counts reaches zero.

Next, all gene–gene pairs found in the literature were contrasted with gene–gene pairs identified by the GMA as co-expressed in at least 50 experiments (Fig. 6). Here there is also a correlation—the more highly connected genes in experimental databases also tend to be studied in the literature, but even as a gene's literature connectivity approaches zero (i.e. little information on gene–gene associations), the co-expression data for the same genes contains more information. Out of 562 933 gene–gene pairs from the GMA, 13 553 (2.4%) could be found in the literature. Conversely, there were 428 472 unique gene–gene pairs in the literature, 21 266 of which were also in the GMA dataset (5% overlap). The former is relatively surprising that so many consistently co-expressed genes do not have publications documenting any relationship between them. The latter is somewhat less surprising since the co-occurrence of genes within publications can be of many different types (e.g. protein–protein interactions, chromosome co-localization, etc.), whereas the GMA data is co-expression only.

4 DISCUSSION AND CONCLUSION

Despite the wide variety of experimental conditions deposited in the GEO database, this study is nonetheless affected and limited

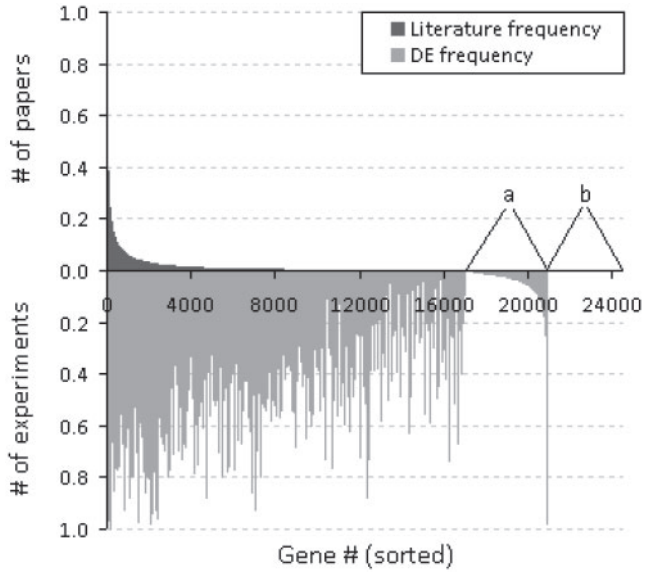

Fig. 5. Comparison of expression data and literature abundance for all human genes. Values are normalized to range between 0 and 1 to enable direct comparison. For each human gene, the frequency of publications per gene (top half) correlates with the frequency of differential expression (bottom half). Arrow 'a' marks where publications per gene reaches zero, yet co-expression data exists for the gene. Arrow 'b' marks where neither co-expression data nor literature is available. 25 390 loci were analyzed, which includes the 24 553 genes used previously plus microRNAs and putative genes.

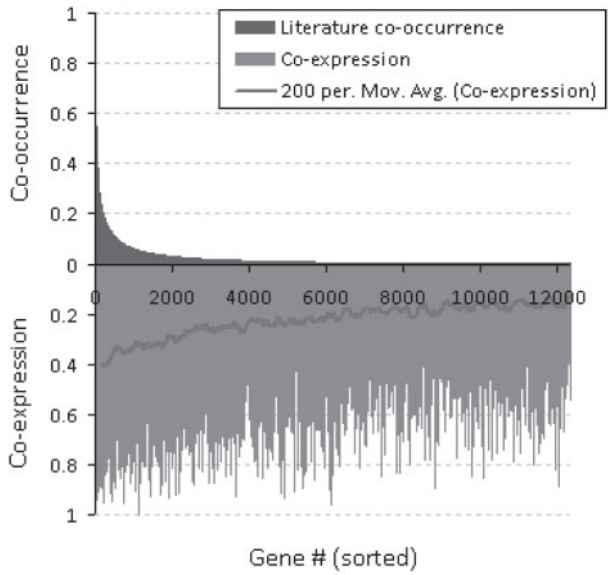

Fig. 6. Frequently co-expressed genes are more likely to co-occur in publications. If co-expression data can be used to predict gene function, then enough data exists to help analyze many genes without much published information.

by the type of experiments conducted (e.g. many cancer samples and cultured cells). For example, some genes were simply not differentially expressed under the conditions studied and it is not yet known how they would behave if they were. Thresholding was used to define differentially expressed genes for mutual information calculations, but any threshold is arbitrary and variations such as weighted mutual information might improve detection of subtle gene–gene co-expression.

Although measuring gene expression with microarray technology has been documented as noisy, raising concerns regarding whether or not highly heterogeneous datasets could be combined in a meaningful manner, this study finds that co-expression patterns tend to recur across datasets and strengthen associations. Figure S3 (Supplementary Material) shows that as the number of co-expressed genes increases, so does the average *P*-value of GO category enrichment analysis.

We find that the genes co-expressed in parallel tend to be much more functionally related than those that are expressed in anti-parallel directions, even though anti-parallel genes tend to have a similar number of significantly enriched GO categories. This seems biologically reasonable considering most genes do not act alone in effecting their biological functions. In prokaryotes, for example, operons link a common promoter to the expression of several genes located one after the other. In eukaryotes operons are rare and regulation is more complex, but the general need for coordinated gene expression remains. Thus, genes whose expression levels rise and fall together, especially across heterogeneous conditions, tend to be those induced or repressed for similar functional reasons. But while a set of genes may be upregulated for a specific biological purpose, the genes downregulated at the same time are likely a mixture of genes that are actively repressed because their function interferes with the newly induced function (e.g. cell migration and adhesion are related activities, yet opposing functions—a cell cannot migrate if it is adhering to another) and those whose function does not conflict with the newly induced function but is merely no longer needed. For example, a differentiating cell may need new surface receptors to create a new function, but some of the existing receptors may degrade not because they interfere, but because there is no need to renew them (and which receptors decay may vary with cell type). This analysis specifically suggests that from an experimental standpoint, gene sets co-expressed with any given gene in both parallel and anti-parallel directions are enriched for significant biological functions, but only gene sets co-expressed in parallel tend to have the *same* function as the gene they are expressed with.

The parallel and anti co-expression patterns shown in this report are probably the most straightforward types, but only reflect general trends, whereby more complex patterns not described here are certainly possible (e.g. downregulated together but not upregulated together). It is this consistency of co-expression under multiple different conditions (experimental, microarray platform, research group, etc) that supports the notion of an informative co-regulatory relationship between genes.

5 ACKNOWLEDGEMENTS

For helpful critiques and suggestions, the author would like to thank Drs. Trey Fondon, Igor Dozmorov, Don Capra and the anonymous peer-reviewers.

Funding: NIH COBRE Junior Investigator award #5P20RR020143-04.

Conflict of Interest: none declared.

REFERENCES

Alexe,G. *et al*. (2005) A robust meta-classification strategy for cancer diagnosis from gene expression data. *Proc IEEE Comput. Syst. Bioinform. Conf*., 322–325.

Bammler,T. *et al*. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods*, **2**, 351–356.

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

Basso,K. *et al*. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet*.,**37**, 382–390.

Cahan,P. *et al.* (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, **401**, 12–18.

Camargo,A.A. *et al*. (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl Acad. Sci. USA*, **98**, 12103–12108.

Choi,H. *et al.* (2007) A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, **8**, 364.

Choi,J.K. *et al.* (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett.*, **565**, 93–100.

Conlon,E.M. *et al*. (2007) Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, **8**, 80.

DeConde,R.P. *et al*. (2006) Combining results of microarray experiments: a rank aggregation approach. *Stat. Appl. Genet. Mol. Biol.*, **5**, Article15.

Dozmorov,I. *et al*. (2004) Hypervariable genes–experimental error or hidden dynamics. *Nucleic Acids Res.,* **32**, e147.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fishel,I. *et al*. (2007) Meta-analysis of gene expression data: a predictor-based approach. *Bioinformatics*, **23**, 1599–1606.

Gauthier,N.P. *et al*. (2008) Cyclebase.org a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res.*, **36** (Database issue), D854–D859.

Ghosh,D. *et al.* (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct. Integr. Genomics*, **3**, 180–188.

Gustin,M.P. *et al*. (2008) Functional meta-analysis of double connectivity in gene co-expression networks in mammals. *Physiol. Genomics.*, **34**, 34–41.

Han,L. and Zhu,J. (2008) Using matrix of thresholding partial correlation coefficients to infer regulatory network. *Biosystems*, **91**, 158–165.

Huminiecki,L. *et al.* (2003) Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics*, **4**, 31.

Jarvinen,A.K. *et al*. (2004) Are data from different gene expression microarray platforms comparable? *Genomics*, **83**, 1164–1168.

Jordan,I.K. *et al*. (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.*, **21**, 2058–2070.

Kato,T. *et al.* (2007) Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. *Cancer Res.*, **67**, 8544–8553.

Khan,J. *et al*. (1999) DNA microarray technology: the anticipated impact on the study of human disease. *Biochim. Biophys. Acta*, **1423**, M17–M28.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lee,H.K. *et al*. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Li,K.C. *et al*. (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA*, **101**, 15561–15666.

Lord,P.W. *et al*. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics*, **19**, 1275–1283.

Luhn,P. *et al*. (2007) Identification of FAKTS as a novel 14-3-3-associated nuclear protein. *Proteins*, **67**, 479–489.

Magwene,P.M. and Kim,J. (2004) Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.*, **5**, R100.

Mariani,T.J. *et al*. (2003) A variable fold change threshold determines significance for expression microarrays. *FASEB J.*, **17**, 321–323.

O'Sullivan,M. *et al.* (2005) Tumor heterogeneity affects the precision of microarray analysis. *Diagn. Mol. Pathol*., **14**, 65–71.

Oldham,M.C. *et al*. (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl Acad. Sci. USA*, **103**, 17973–17978.

Pena-Castillo,L. and Hughes,T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics*, **176**, 7–14.

Pritchard,C. *et al.* (2006) The contributions of normal variation and genetic background to mammalian gene expression. *Genome Biol.*, **7**, R26.

Pritchard,C.C. *et al.* (2001) Project normal: defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci. USA*, **98**, 13266–13271.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95*), pp. 448–453.

Rhodes,D.R. *et al*. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.

Rhodes,D.R. *et al*. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.

Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Spellman,P.T. *et al* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell*, **9**, 3273–3297.

Stevens,J.R. and Doerge,R.W. (2005) Combining Affymetrix microarray results. *BMC Bioinformatics*, **6**, 57.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Suarez-Farinas,M. and Magnasco,M.O. (2007) Comparing microarray studies. *Meth. Mol. Biol.*, **377**, 139–152.

Ucar,D. *et al*. (2007) Construction of a reference gene association network from multiple profiling data: application to data analysis. *Bioinformatics*, **23**, 2716–2724.

van Noort,V. *et al*. (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*., **5**, 280–284.

Wang,J. *et al*. (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics*, **20**, 3166–3178.

Wren,J.D. (2004) Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, **5**, 145.

Wren,J.D. *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**, 389–398.

Wren,J.D. and Conway,T. (2006) Meta-analysis of published transcriptional and translational fold changes reveals a preference for low-fold inductions. *OMICS*, **10**, 15–27.

Wren,J.D. and Garner,H.R. (2002) Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inf. Med.*, **41**, 426–434.

Wren,J.D. and Garner,H.R. (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, **20**, 191–198.

Yan,X. *et al.* (2007) A graph-based approach to systematically reconstruct human transcriptional regulatory modules. *Bioinformatics*, **23**, i577–i586.

Yang,X. and Sun,X. (2007) Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers. *BMC Bioinformatics*, **8**, 118.

Yip,A.M. and Horvath,S. (2007) Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, **8**, 22.

Yoon,S. *et al*. (2006) Large scale data mining approach for gene-specific standardization of microarray gene expression data. *Bioinformatics*, **22**, 2898–2904.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article17.